

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
)
NAKAMURA et al.)
)
Application Number: To Be Assigned)
)
Filed: Concurrently Herewith)
)
For: CLUSTERING DISK CONTROLLER, ITS DISK)
CONTROL UNIT AND LOAD BALANCING METHOD)
OF THE UNIT)

jc872 U.S. PTO
10/090767
03/06/02



Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

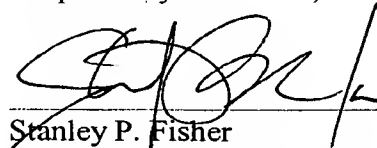
**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of January 10, 2002, the filing date of the corresponding Japanese patent application 2002-002936.

The certified copy of corresponding Japanese patent application 2002-002936 is being submitted herewith. Acknowledgment of receipt of the certified copy is respectfully requested in due course.

Respectfully submitted,



Stanley P. Fisher
Registration Number 24,344

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
March 6, 2002

JUAN CARLOS A. MARQUEZ
Registration No. 34,072

日 本 国 特 許 庁
JAPAN PATENT OFFICE

Jc872 U.S. PTO
10/090767
03/06/02

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application: 2002年 1月10日

出 願 番 号

Application Number: 特願2002-002936

[ST.10/C]:

[JP2002-002936]

出 願 人

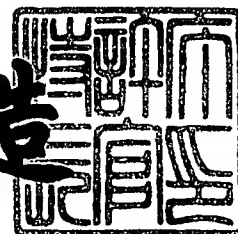
Applicant(s): 株式会社日立製作所

CERTIFIED COPY OF
PRIORITY DOCUMENT

2002年 2月15日

特 許 庁 長 官
Commissioner,
Japan Patent Office

及 川 耕 造



出証番号 出証特2002-3006902

【書類名】 特許願

【整理番号】 H01013571A

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 3/06

【発明者】

【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所中央研究所内

【氏名】 中村 崇仁

【発明者】

【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所中央研究所内

【氏名】 藤本 和久

【発明者】

【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所中央研究所内

【氏名】 金井 宏樹

【発明者】

【住所又は居所】 神奈川県小田原市中里 3 2 2 番地 2 号 株式会社日立製作所 R A I D システム事業部内

【氏名】 吉田 晃

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社 日立製作所

【代理人】

【識別番号】 100075096

【弁理士】

【氏名又は名称】 作田 康夫

【電話番号】 03-3212-1111

【手数料の表示】

【予納台帳番号】 013088

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 クラスタ型ディスク制御装置および負荷分散方法

【特許請求の範囲】

【請求項 1】

複数台のディスク制御装置と該複数のディスク制御装置を接続する接続手段とを備えたクラスタ型ディスク制御装置において、前記ディスク制御装置に備えられたチャンネル制御部と、前記クラスタ型ディスク制御装置内に設けられたスイッチとを有し、該スイッチは前記チャンネル制御部及びホストコンピュータと接続され、前記スイッチは、前記ホストコンピュータから指定されたアクセス要求先である宛先チャンネル制御部と該アクセス要求を実際に転送するチャンネル制御部との対応情報を保持するデータテーブルを備えたことを特徴とするクラスタ型ディスク制御装置。

【請求項 2】

複数台のディスク制御装置と該複数のディスク制御装置間の通信手段を備えたクラスタ型ディスク制御装置において、前記ディスク制御装置に備えられたチャンネル制御部と、前記クラスタ型ディスク制御装置内に設けられたスイッチとを有し、該スイッチは前記チャンネル制御部と接続され、前記スイッチはホストコンピュータと接続され、前記スイッチは、前記ホストコンピュータからアクセス要求のあったチャンネル制御部とは別の前記チャンネル制御部に転送するか否かの情報を保持する分配テーブルを有することを特徴とするクラスタ型ディスク制御装置。

【請求項 3】

請求項 1 に記載のクラスタ型ディスク制御装置において、前記ホストコンピュータからのアクセス要求の転送先として複数のチャンネル制御部が指定可能であり、前記データテーブルには、前記複数のチャンネル制御部の各々が転送先として選ばれる確率が格納されることを特徴とするクラスタ型ディスク制御装置。

【請求項 4】

請求項 1 に記載のクラスタ型ディスク制御装置において、ディスク制御装置内情報を管理するサービスプロセッサ（SVP）を有し、該 SVP は前記データテーブルを変更することを特徴とするディスク制御装置。

【請求項 5】

請求項 1 に記載のクラスタ型ディスク制御装置において、前記ディスク制御装置は、ホストコンピュータからのアクセス要求を受信したチャネル制御部が該当アクセス要求を処理のどの途中段階まで実行するかを有する代行レベルテーブルを備えることを特徴とするディスク制御装置。

【請求項 6】

複数台のディスク制御装置と、該複数のディスク制御装置を接続する接続手段と、チャネル制御部と、ホストコンピュータからのアクセス要求をチャネル制御部へ転送するためのデータテーブルを備えたスイッチとを有するディスクサブシステムの制御方法において、

ホストコンピュータからのアクセス要求を前記データテーブルに基づき所定のチャネル制御部へ転送し、前記アクセス要求が転送されたチャネル制御部がアクセス要求を処理し、前記アクセス要求の返答として、前記ホストコンピュータからアクセス要求先として指定された宛先チャネル制御部が返答の発行元であるとするデータをホストコンピュータに対して送信することを特徴とするディスクサブシステムの制御方法。

【請求項 7】

請求項 6 に記載のディスクサブシステムの制御方法において、前記アクセス要求が転送されたチャネル制御部が前記ホストコンピュータからのアクセス要求の宛先とは異なる場合、前記要求を受信したチャネル制御部が途中段階まで前記要求を処理し、前記要求の宛先チャネル制御部に途中段階までの処理終了を通知し、処理の残りを前記要求宛先チャネル制御部が実行することを特徴とするディスクサブシステムの制御方法。

【請求項 8】

請求項 6 に記載のディスクサブシステムの制御方法において、前記ディスクサブシステムは、ホストコンピュータからのアクセス要求を受信したチャネル制御部が該当アクセス要求を処理のどの途中段階まで実行するかを有する代行レベルテーブルを備え、

前記チャネル制御部は、前記スイッチから転送されてきたアクセス要求を、前記

代行レベルテーブルに記載された情報に基づき中途段階まで処理し、前記ホストコンピュータが実際に指定したアクセス要求先である宛先チャネル制御部に途中段階までの処理終了を通知し、前記中途段階まで処理されたアクセス要求の処理の残りを前記要求宛先チャネル制御部が実行することを特徴とするディスクサブシステムの制御方法。

【請求項 9】

請求項 6 に記載のディスクサブシステムの制御方法において、前記ディスクサブシステムは、ディスク制御装置内情報を管理するサービスプロセッサ（SVP）を備え、前記 SVP は、各チャネル制御部の負荷情報を参照し、負荷が高いチャネル制御部を宛先としたホストコンピュータからのアクセス要求を負荷が低いチャネル制御部に転送するように前記データテーブルを変更することを特徴とするディスクサブシステムの制御方法。

【請求項 10】

請求項 9 に記載のディスクサブシステムの制御方法において、前記 SVP は、各チャネル制御部の障害情報を参照し、障害下にあるチャネル制御部を宛先としたホストコンピュータからのアクセス要求を正常なチャネル制御部に転送するように、前記テーブルを変更することを特徴としたディスクサブシステムの制御方法。

【請求項 11】

請求項 9 に記載のディスクサブシステムの制御方法において、前記 SVP は、各チャネル制御部の負荷情報を参照し、負荷が低いチャネル制御部に対する処理の実行段階を上げるように前記代行レベルテーブルを変更することを特徴としたディスクサブシステムの制御方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、データを複数の磁気ディスク装置に格納するディスクサブシステムに関し、特にクラスタ型ディスク制御装置とその負荷分散方法に関する。

【0002】

【従来の技術】

多数台の磁気ディスク装置（以下ドライブと呼ぶ）に対するデータの格納および読み出しを行うディスク制御装置（以下DKCと呼ぶ）があり、ドライブとDKCとをあわせてディスクサブシステムと総称されている。

ディスクサブシステムに対する要求の1つとして記憶容量の増大および管理コストの削減とがある。1台のDKCで管理可能なドライブの容量には限界があるため、複数のディスクサブシステムを用意して記憶容量を増加する。すると、その管理に必要となる管理コストも同様に増大するのである。そこで、従来サーバ毎に接続されこの結果分散配置されていたディスクサブシステムの集中化を図るべくストレージエリアネットワーク（以下SANと呼ぶ）が注目されている。

図2には、SAN環境におけるディスクサブシステムの典型例を示す。複数台のディスクサブシステム1がSANスイッチ39を介してホストコンピュータ0に接続される。1つのディスクサブシステムは1台のディスク制御装置10のみから構成されていて、チャンネル2を介してSANスイッチ39と接続されている。論理ディスク7は、ホストコンピュータ0が認識する記憶領域である。ホストコンピュータ0は、SANスイッチ39とチャンネル2を介して論理ディスク7の特定のアドレスに対してデータの参照、更新要求を指示する。チャンネル2としては、ファイバチャンネル、SCSIなどがある。ディスク制御装置10と複数台のドライブ17は、ドライブIF16で接続される。ドライブIF16には、ファイバチャンネル、SCSIなどが用いられる。DKC10は、大きくは、チャンネルの制御を行うチャンネル制御部11、ドライブの制御を行うディスク制御部14、DKCの制御情報3を格納する制御メモリ部12、キャッシュデータ5を保持するキャッシュメモリ部13、さらに、各構成部品を相互に接続する結合機構15から構成される。結合機構15は、バス、相互結合網などが用いられる。ディスク制御装置10は、ホストコンピュータ0の指示に従い、データの参照、更新処理を行う。

一方、SAN環境においては、ホストコンピュータは、アクセス対象とするデータがどのディスクサブシステムに存在しているかを知らなければデータにアクセスできない。したがって、ユーザがデータの所在を管理せねばならないという問

題がある。特開平 2 0 0 0 - 9 9 2 8 1 号公報には、従来 1 台の D K C から構成していたディスクサブシステムを、S A N スイッチを介さずに複数台の D K C でクラスタ構成して記憶容量と接続チャネル数を増大したディスクサブシステムが開示されている。クラスタ型ディスクサブシステムは、1 台のディスクサブシステムとして管理が可能なので管理コストも削減できる。

図 3 には、クラスタ型ディスクサブシステムの構成例を示す。図 3 のディスクサブシステムでは、複数の D K C 1 0 間にディスク制御装置間接続手段 2 0 を設けることにより D K C 間での相互データアクセスを可能としている。これにより複数のディスク制御装置間でのデータの共有が可能になる。

【0 0 0 3】

【発明が解決しようとする課題】

図 3 に示すような従来のクラスタ構成のディスクサブシステムにおいては、内部の D K C 間で負荷に偏りが生じた場合に負荷を分散することができず、内部 D K C 間で同等の性能を引き出すことができない。ホストコンピュータからのアクセス要求が特定 D K C に集中した場合、別の D K C の稼働率が低い場合においても、ホストコンピュータが指定した要求先 D K C が処理を実行するので、要求先として特定 D K C によってしか処理が実行されないため負荷に偏りが生じるのである。

【0 0 0 4】

また、チャネル制御部が障害に陥った場合には、そのチャネル制御部に接続されているチャネルは使用不可能になってしまう。ホストコンピュータが指定したチャネルのみを経由してディスク制御装置にアクセスするためである。

【0 0 0 5】

【課題を解決するための手段】

クラスタ構成のディスクサブシステムの内部にチャネルー D K C 間スイッチを設け、更にチャネルー D K C 間スイッチにホストコンピュータからのアクセス要求をどのチャネル制御部に転送するか情報を保持する分配テーブルを設け、分配テーブルに基づいて要求を転送するようにした。

【0 0 0 6】

更に、装置情報を管理するサービスプロセッサ（SVP）を設けて分配テーブルを書き換えることにより、装置負荷や障害状況に基づきチャネル-DKC間スイッチがホストコンピュータのアクセス要求を適当なチャネル制御部に転送できるようにした。

【0007】

【発明の実施の形態】

以下、図面を用いて、発明の詳細を説明する。

【0008】

はじめに、図1を用いて、本発明に係るディスク制御装置について説明する。図1は、本発明に係わるディスク制御装置の概要を示すブロック図の一例である。ディスクサブシステム1は複数台のDKC10とチャネル-DKC間スイッチ30と複数のドライブにより構成する。ディスクサブシステム1は、複数のチャネル2を介して、複数のホストコンピュータ0に接続する。チャネル-DKC間スイッチ30はホストコンピュータ0からの要求を分配テーブル31に基づいて該当DKC10に転送する。各DKC10は、各DKC10の間を接続するディスク制御装置間接続手段20を介して、他のDKCに接続したドライブ17に格納した格納データ4を参照、更新できる。本発明では、特に、ホストコンピュータ0からのアクセス要求をチャネル-DKC間スイッチ30が受信し、アクセス要求の宛先と分配テーブル31に基づいて、該当アクセス要求を該当DKC10のチャネル制御部11に転送することに特徴がある。

【0009】

以下、詳細に説明するが先立って、用いる用語の定義を行う。本発明では、ホストコンピュータ0からのアクセス要求の宛先となるDKC10のチャネル制御部11を宛先チャネル制御部と呼ぶ。チャネル-DKC間スイッチ30の転送先変更により該当アクセス要求を受信したチャネル制御部11を要求受信チャネル制御部または代行チャネル制御部と呼ぶ。代行チャネル制御部が、該当アクセス要求に対する処理を行うことを代行と呼び、その処理を代行処理と呼ぶ。また、代行チャネル制御部が代行を途中段階で中断し、他のチャネル制御部が継続して処理を行うことを引継ぎと呼び、その処理を引継ぎ処理と呼ぶ。また、代行チャ

ネル制御部がアクセス要求に対する処理を全て行うことを完全代行と呼び、その処理を完全代行処理と呼ぶ。

【 0 0 1 0 】

図 1 に示したディスクサブシステム 1 は、大きくは、複数台の D K C 1 0 と、ドライブ 1 7、チャネルー D K C 間スイッチ 3 0 から構成する。D K C 1 0 の詳細は 2 台のみ詳細に示しているが、各 D K C 1 0 の各々の構成は同一である。D K C 1 0 は、大きくは、チャネルの制御を行うチャネル制御部 1 1、ドライブの制御を行うディスク制御部 1 4、D K C 1 0 の制御情報 3 を格納する制御メモリ部 1 2、ドライブ 1 7 のデータを一時的にキャッシュデータ 5 として保持するキャッシュメモリ部 1 3、さらに、各構成部品を相互に接続する結合機構 1 5 から構成する。チャネル制御部 1 1 は、単独のチャネル制御部に複数のチャネルを持つ構成でもよい。しかし以降の説明では、簡単化のため、1 つのチャネル制御部 1 1 は 1 つのチャネル 2 のみを持つ場合を示す。このため、後述するチャネル制御部の識別子とチャネルの識別子は一致している。制御メモリ部 1 2 は、キャッシュメモリ部 1 3 内データのディレクトリ情報も格納されており、各キャッシュメモリ部 1 3 に要求データが存在するかどうかを確認できる。また、制御メモリ部 1 2 は装置構成情報も格納しており、アクセス要求先のアドレスにより、アクセス要求先のディスク制御装置番号とドライブ番号を識別できる。図示はしていないが、チャネル制御部 1 1 やディスク制御部 1 4 は、制御用のプロセサを備え、プロセサ上で処理プログラムが動作する。チャネルー D K C 間スイッチ 3 0 は、分配テーブル 3 1 を備える。

図 4 は、図 1 におけるチャネルー D K C 間スイッチ 3 0 の構成とチャネルー D K C 間スイッチに備えた分配テーブル 3 1 に格納する情報の一例を示したブロック図である。チャネルー D K C 間スイッチは、ホストコンピュータ 0 とチャネルを介して接続するホストコンピュータ側入出力部 3 0 1 と、D K C に接続する D K C 側入出力部 3 0 2 を備える。ホストコンピュータ側入出力部 3 0 1 は複数のホスト側ポート 3 0 1 1 を、D K C 側入出力部 3 0 2 は複数の D K C 側ポート 3 0 2 1 を備える。これらが相互結合網 3 0 3 を介して接続されており、任意のホスト側ポート 3 0 1 1 と任意の D K C 側ポート 3 0 2 1 でデータ転送が可能とな

る。一般にポートはそれぞれに識別子をもつ。例えばインターネットプロトコルならばMACアドレス、ファイバチャネルプロトコルならばWWNである。ここでは一般化のため、番号として表記し、各々のチャネル制御部の識別子をチャネル制御部番号、ホストコンピュータの識別子をホストコンピュータ番号などとする。チャネル-DKC間スイッチ30は、DKC側から来たデータに関してはホストコンピュータポート番号変換テーブル32に基づき該当するDKC側ポートが該当するホスト側ポートに転送する。ホストコンピュータ側から来たデータ（アクセス要求）に関しては分配テーブル31に基づき転送を行う。

【0011】

分配テーブル31は、ホストコンピュータからのアクセス要求の宛先チャネル制御部とその要求を転送するDKC側ポート番号の対応情報を保持する。本実施例では、ホストコンピュータからのアクセス要求を宛先チャネル制御部に対応する転送先DKC側ポート番号の要素が0の場合は、その転送先DKC側ポート3021には該当する宛先の要求を転送せず、それ以外の場合には、宛先が該当チャネル制御部番号の要求を転送するDKC側ポート3021の候補となる。ここでは、要素上段のように転送先候補のものを1とするものや、要素下段のようにそれぞれが選択される確率を示すものでもよい。前者の方法では、複数の転送先が1の場合は、ラウンドロビンなどの方法で転送先のチャネル制御部を選択する。後者の方法については、図25を用いて後述する。本実施例図4で示した分配テーブル31の要素は、チャネル制御部番号3のチャネル制御部は障害下にあり、チャネル制御部番号1のチャネル制御部は高負荷状態にある場合の例である。分配テーブル31の宛先チャネル制御部番号が3の列は、障害下にあるチャネル制御部番号3のチャネル制御部に代わってチャネル制御部番号5のチャネル制御部がアクセス要求を処理することを示している。また、宛先チャネル制御部番号が1の列は、チャネル制御部番号4が高負荷状態にあるチャネル制御部番号1の宛先の要求を半分の割合で処理することを示している。分配テーブル31は1つの宛先に対して複数の転送先を設定可能な点が特徴である。分配テーブル31を用いることにより、要求を本来の宛先チャネル制御部から他のチャネル制御部に割り振ることが可能になりチャネル制御部11の負荷分散や障害処理が可能にな

る。

【 0 0 1 2 】

図 5 は、図 1 におけるチャネル制御部 1 1 に備えられた代行レベルテーブル 1 1 3 に格納する情報の一例を示したブロック図である。チャネル制御部は制御プロセッサ 1 1 0 を有し、制御プロセッサ部 1 1 0 はホスト要求処理部 1 1 1 とモニタ機構部 1 1 2 を備える。ホスト要求処理部 1 1 1 により、チャネル制御部 1 1 が受領したホストコンピュータのアクセス要求を処理する。またモニタ機構部 1 1 2 により、SVP 4 0 にチャネル制御部 1 1 の状態を報告する。本実施例では、チャネル制御部 1 1 に代行レベルテーブル 1 1 3 を備えているが、該代行レベルテーブルは、制御メモリ部 1 2 に備えても良い。

【 0 0 1 3 】

代行レベルテーブル 1 1 3 は、宛先チャネル制御部に代わりに、要求を受信したチャネル制御部がどの段階まで処理を行うかの段階を示す代行レベル情報を保持する。例えば、チャネル制御部 1 1 のキャッシュミス時のリード要求は、コマンド解析（段階 1）、キャッシュヒット判定（段階 2）、キャッシュにドライブのデータを格納するステージング処理をディスク制御部に要求（段階 3）、ステージング処理終了判定（段階 4）、データ転送（段階 5）となるとする。本実施例の代行レベルテーブル 1 1 3 では、該当チャネル制御部がチャネル制御部番号 4 に対するリード要求に関しては段階 2、つまり、キャッシュヒット判定までの処理を行い、段階 3 以降の処理はチャネル制御部番号 4 のチャネル制御部が行う。また、チャネル制御部番号 6 に対するアクセス要求に関しては該当チャネル制御部が段階 1 から段階 5 までの全ての処理を行う（完全代行）。本実施例では、リード／ライトアクセスに対する代行レベルを分けているが、リードおよびライトアクセスに対して同一の代行レベルとなるような代行レベルテーブルとしても良い。

【 0 0 1 4 】

図 6 は、図 1 における制御メモリ 1 2 に備えた代行処理情報 1 2 4 に格納する情報の一例を示したブロック図である。制御メモリ 1 2 は、制御情報 1 2 1、ディレクトリ情報 1 2 2、装置構成情報 1 2 3、代行処理情報 1 2 4 を有する。制

御情報 1 2 1 は、装置制御に用いる情報が格納されており、例えば、チャンネル制御部 1 1 がディスク制御部 1 4 にデータのステージング処理を依頼するなど用いる。ディレクトリ情報 1 2 2 は、データの存在する装置の対応を示す。装置構成情報 1 2 3 は、各部の存在や容量などの情報を保持している。次に、代行処理情報 1 2 4 について説明する。

【 0 0 1 5 】

代行処理情報 1 2 4 は、処理の引継ぎに関する情報であり、ジョブ番号 1 2 4 1、要求ホスト 1 2 4 2、要求宛先 1 2 4 3、コマンド内容 1 2 4 4、処理段階 1 2 4 5、データ存在アドレス 1 2 4 6 を有する。ジョブ番号 1 2 4 1 は、アクセス要求処理に対する識別番号で、ディスク制御部に対するデータ転送を依頼などに用いる。要求ホスト 1 2 4 2 はアクセス要求をしたホストコンピュータ 0 の識別子である。要求宛先 1 2 4 3 は該当ホストコンピュータ 0 が発行した要求の本来の宛先チャンネル制御部識別子である。要求宛先 1 2 4 3 を参照することにより、該当要求宛先以外のチャンネル制御部が該当要求宛先のチャンネル制御部から要求に対する応答を返したような情報を応答に記載することが可能になる。コマンド内容 1 2 4 4 は、要求を受領したチャンネル制御部行った要求の解析結果が、リード、ライト等どのコマンドであるかをあらわす情報である。処理段階 1 2 4 5 は、要求が受領したチャンネル制御部がどの段階までの処理を行ったかを示す情報である。処理段階 1 2 4 5 には、キャッシュヒット／ミス等の処理の進行に関する情報も含む。データ存在アドレス 1 2 4 6 は代行した処理段階で得られたデータが存在するアドレスを示す。

【 0 0 1 6 】

本実施例では、制御メモリ部 1 2 に代行処理情報 1 2 4 を格納しているが、代行処理情報 1 2 4 をチャンネル制御部間でメッセージとして伝達してもよい。

【 0 0 1 7 】

図 2 0 は、図 1 における S V P 4 0 の詳細である。S V P 4 0 は制御プロセッサ 4 0 0、負荷情報テーブル 4 0 1、チャンネル制御部－ポート番号変換テーブル 4 0 2、ポート番号－チャンネル制御部変換テーブル 4 0 3、装置管理インタフェース部 4 0 4 を有する。制御プロセッサ 4 0 0 は障害監視部 4 0 0 1、負荷監視

部 4 0 0 2、テーブル指示部 4 0 0 3、ローカルメモリ 4 0 0 4 を有する。障害監視部 4 0 0 1 は、ディスクサブシステム 1 各部の定期的なアクセスや各部からの報告を受ける機能を行う部位であり、ディスクサブシステム 1 各部の障害情報を管理する。負荷監視部 4 0 0 2 は、ディスクサブシステム 1 各部の負荷率を各部からの報告を受けることにより計測する機能で、各部の負荷状況を管理する。テーブル指示部 4 0 0 3 は、分配テーブル 3 1 または代行レベルテーブル 1 1 3 を変更する機能を持つ。ローカルメモリ 4 0 0 4 は負荷率順にソートした結果を収容するなど、一連の手続きで必要となる一時的な情報を格納する。装置管理インタフェース部 4 0 4 は、装置管理者が装置設定を行う場合や装置状態を確認する場合の入出力インタフェースである。負荷情報テーブル 4 0 1 は、負荷監視部 4 0 0 2 により得られた各チャネル制御部の負荷率を格納する。負荷情報テーブル 4 0 1 を参照し、各チャネル制御部を負荷率順でソートすることなどが可能となる。チャネル制御部—ポート番号変換テーブル 4 0 2 とポート番号—チャネル制御部変換テーブル 4 0 3 は、チャネル制御部 1 1 番号と D K C 側ポート 3 0 2 1 番号の対応情報を保持する。これらにより、後述する S V P 4 0 による分配テーブル 3 1 の更新が、チャネル制御部 1 1 番号と D K C 側ポート 3 0 2 1 番号の関係がどのような場合であっても可能となる。

【 0 0 1 8 】

S V P 4 0 が分配テーブル 3 1 の参照や更新を行う場合、チャネル制御部—ポート番号変換テーブル 4 0 2 やポート番号—チャネル制御部変換テーブル 4 0 3 に照らし合わせ、実際のチャネル制御部番号に対応する D K C 側ポート番号を知る。以降では簡単化のため、S V P はチャネル制御部番号に対応する D K C 側ポート番号を予め知っているものとして「該当転送先チャネル制御部の行」等の記述を用いる。

【 0 0 1 9 】

本実施例では、負荷情報テーブル 4 0 1、チャネル制御部—ポート番号変換テーブル 4 0 2、ポート番号—チャネル制御部変換テーブル 4 0 3 を S V P 4 0 に格納したが、制御メモリ部 1 2 やチャネル—D K C 間スイッチ 3 0 に格納しても良い。

【 0 0 2 0 】

次に流れ図を用いてチャネル制御部 1 1 および S V P 4 0 の制御方法を説明する。この制御方法は、チャネル制御部が負荷の高い状態や障害状態にある場合、チャネルー D K C 間スイッチ 3 0 によりアクセス要求が該当チャネル制御部とは別のチャネル制御部に転送して負荷分散およびフェイルオーバを可能にするものである。本発明の制御方法の形態は大きく 2 つある。アクセス要求を受信したチャネル制御部が処理をすべて行う完全代行の形態と、アクセス要求を受信したチャネル制御部が途中段階まで処理を行い（代行）その後該当アクセス要求の本来の宛先であるチャネル制御部が該当処理を継続し該当アクセス要求処理を完了する（引継ぎ）代行引継ぎの形態である。この 2 つの形態ごとに、リード・ライト要求に対するチャネル制御部 1 1 の制御方法、チャネル制御部障害発生時および障害回復時の S V P 4 0 の制御方法、チャネル制御部間の負荷に偏りが生じた場合の S V P 4 0 の制御方法を説明する。なお、ここでは分配テーブル 3 1 の実施形態として、分配テーブルの要素が転送先候補とするか否かの 2 値とした場合と転送先とする確率を示した場合とがある。S V P 4 0 のそれぞれの制御方法では分配テーブルの各実施形態についての説明を行う。

【 0 0 2 1 】

まず、完全代行の制御方法を説明する。図 7 はリードアクセス要求に、図 8 はライトアクセス要求に対するチャネル制御部 1 1 の制御方法を示す流れ図である。この方法で制御するチャネル制御部を備える装置で、S V P 4 0 が分配テーブル 3 1 を変更することによりチャネル制御部の障害を回避する制御方法を流れ図で示したものが図 9 である。同様に該当チャネル制御部が障害から回復した場合の S V P の制御方法を流れ図で示したものが図 1 0 である。また、チャネル制御部間の負荷に偏りが生じた場合に、分配テーブル 3 1 を変更し高い負荷状態にあるチャネル制御部に転送するアクセス要求の割合を減らすことにより負荷分散する S V P の制御方法を図 1 1 に示す。図 9、図 1 0、図 1 1 はいずれも分配テーブルの要素が転送先候補とするか否かの 2 値で示す場合の実施例であり、分配テーブルの要素を転送先とする確率を示す場合の対応する実施例は図 1 4、図 1 3、図 1 2 に示す。

【 0 0 2 2 】

図 7 は、リード要求受領時の処理の一例を示す流れ図である。リード要求を受領すると、受信要求から、本来の要求宛先のチャネル制御部とコマンドとアクセス先アドレスを解析し、リードアクセスであることを認識する（ステップ 1）。アクセス先アドレスは、制御メモリ部 1 2 の装置構成情報 1 2 3 を参照することで、アクセス要求先のディスク制御装置番号とドライブ番号を識別できる。次に、ステップ 1 で識別した当該 D K C のキャッシュに対してキャッシュヒットミス判定を行う（ステップ 2）。制御メモリ部 1 2 のディレクトリ情報 1 2 2 を参照することで、アクセス先データがキャッシュに保持されているかを識別可能である。キャッシュに保持しているか判定し（ステップ 3）、キャッシュに保持していないキャッシュミスの場合は、当該 D K C のディスク制御部に対して当該データのドライブからキャッシュへの転送依頼を行う（ステップ 5）。通常この処理はステージング処理と呼ばれる。この場合転送終了までリード処理を中断し（ステップ 6）、ステージング終了後、再びリード処理を継続することになる。また、転送先のキャッシュアドレスは、キャッシュの空きリストなど一般的な方法で管理、取得すればよいが、転送先アドレスをディレクトリ情報 1 2 2 を更新することで登録する必要がある。ステップ 3 でヒット判定の場合、または、ステップ 7 でステージング処理が終了した場合は、ホストコンピュータに対して当該データを転送する（ステップ 4）。ステップ 4 の際、処理を行ったチャネル制御部がホストコンピュータのアクセス要求の宛先と異なっても、該当チャネル制御部と異なる本来の宛先チャネル制御部の識別子を応答データの発信元として応答データに付与し、本来の宛先チャネル制御部が応答したとするようにデータを転送する。この点が本実施例の特徴である。

図 8 は、ライト要求受信時の処理の一例を示す流れ図である。ライト要求を受領すると、受信要求から、本来の要求宛先のチャネル制御部とコマンドとアクセス先アドレスを解析し、ライトコマンドであると認識する（ステップ 1）。アクセス先アドレスは、制御メモリ部 1 2 の装置構成情報 1 2 3 を参照することで、アクセス要求先のディスク制御装置番号とドライブ番号を識別できる。次に、ステップ 1 で識別した当該 D K C のキャッシュに対してキャッシュヒットミス判定

を行う（ステップ2）。制御メモリ部12のディレクトリ情報122を参照することで、アクセス先データがキャッシュに保持されているかを識別可能である。キャッシュに保持していないキャッシュミスの場合は、当該DKCのディスク制御部対して当該データのドライブからキャッシュへの転送依頼を行う（ステップ6）。通常この処理はステージング処理と呼ばれる。この場合転送終了までライト処理を中断し（ステップ7）、ステージング終了後、再びライト処理を継続することになる。また、転送先のキャッシュアドレスは、キャッシュの空きリストなど一般的な方法で管理、取得すればよいが、転送先アドレスをディレクトリ情報122を更新することで登録する必要がある。ステップ3でヒット判定の場合、または、ステップ7でステージング処理が終了した場合は、当該DKCのキャッシュに対して当該データの更新を行う（ステップ4）。更新終了後、ホストコンピュータに対してライト処理の完了報告を行う（ステップ5）。ステップ5の際、処理を行ったチャネル制御部がホストコンピュータのアクセス要求の宛先と異なっているとしても、本来の宛先チャネル制御部が応答したとする完了報告をする。この点が本実施例の特徴である。

図9は、あるチャネル制御部11が障害に陥った場合に、該当チャネル制御部をフェイルオーバーするSVP40の処理の一例を示す流れ図である。SVPが更新する分配テーブル31の要素は、転送先候補とするか否かの2値の場合の説明をする。ここでは障害下にあるチャネル制御部を障害チャネル制御部と呼ぶ。SVP40の障害監視部4001により、チャネル制御部が障害となったことを認識する（ステップ1）。次に負荷監視部4002により得られた正常なチャネル制御部の負荷率を障害情報テーブル401より参照し、最低負荷率のチャネル制御部を見つける（ステップ2）。その後、分配テーブル31を参照し要求転送先チャネル制御部が障害チャネル制御部の行を順次チェックする（ステップ3）。該当要素が1、すなわちチャネルーDKC間スイッチ30が障害チャネル制御部に要求を転送することになっていた場合は、該当要素を0として障害チャネル制御部に要求を転送しないようにし、また、該当要素の列の要求転送先チャネル制御部が該当最低負荷率のチャネル制御部の要素を1にして要求転送先チャネル制御部が無い状態を防ぐ（ステップ4）。ステップ3、ステップ4を該当行全体に

対してチェック終了するまで続ける。

【 0 0 2 3 】

図 1 0 は、障害下にあったチャネル制御部が障害から回復した場合の S V P 4 0 が分配テーブル 3 1 を更新する処理の一例を示す流れ図である。S V P が更新する分配テーブル 3 1 の要素は転送先候補とするか否かの 2 値の場合の説明をする。S V P 4 0 の障害監視部 4 0 0 1 により、障害下にあったチャネル制御部が障害から回復したことを確認する（ステップ 1）。それを受けて、分配テーブル 3 1 の要求宛先チャネル制御部の列と要求転送先チャネル制御部の行が共に該当チャネル制御部である要素を 1 にして、該当チャネル制御部に要求が転送されるようにする（ステップ 2）。該当チャネル制御部のフェイルオーバーしていたチャネル制御部はこの処理以降の負荷分散を目的とした分配テーブル 3 1 の変更により、順次、代行を解除される。

【 0 0 2 4 】

図 1 1 は、各チャネル制御部 1 1 の負荷に偏りが生じた場合における S V P 4 0 の処理の一例を示す流れ図である。S V P 4 0 の負荷監視部 4 0 0 2 により得られた負荷情報テーブル 4 0 1 から各チャネル制御部の負荷率に偏りがあることを確認する（ステップ 1）。負荷率の偏りは、例えば、最高負荷率と最低負荷率の差が閾値を超えた場合で定義できる。次に、チャネル制御部を負荷率昇順でソートする（ステップ 2）。その結果により最高負荷率のチャネル制御部の負荷を下げるように分配テーブル 3 1 を更新する。分配テーブルで要求宛先チャネル制御部が該当最高負荷率のチャネル制御部の列をステップ 2 のソート順にチェックしていく（ステップ 3）。該当要素が 0 ならば、該当要素を 1 として負荷分散先として登録する（ステップ 6）。該当要素が 1 ならば次の要素をチェックする。全要素チェック後またはステップ 6 終了後は、分配テーブルで要求転送先チャネル制御部が該当最高負荷率のチャネル制御部である行をステップ 2 のソート順にチェックしていく（ステップ 7）。該当要素が 1 ならば、該当要素を 0 として該当最高負荷率のチャネル制御部を負荷分散先から削除する（ステップ 1 0）。該当要素が 0 ならば次の要素をチェックする。全要素チェック後またはステップ 1 0 終了後は、この処理を終了とする。

【 0 0 2 5 】

図 1 2 は、各チャネル制御部 1 1 の負荷に偏りが生じた場合における S V P 4 0 の処理の一例を示す流れ図である。S V P が更新する分配テーブル 3 1 の要素は転送先とする確率を示したものである場合の説明をする。S V P 4 0 の負荷監視部 4 0 0 2 により得られた負荷情報テーブル 4 0 1 から各チャネル制御部の負荷率に偏りがあることを確認する（ステップ 1）。次に正常チャネル制御部の負荷率が最低であるものと最高であるものを選択する（ステップ 2）。分配テーブル 3 1 中の最高負荷率のチャネル制御部が転送先ポートの行について全ての Δ 以上の要素から Δ を減算する（ステップ 3）。この際、障害下にあるチャネル制御部に該当する要素については除外する。さらに、分配テーブル 3 1 中の最低負荷率のチャネル制御部が転送先ポートの行についてステップ 3 にて減算した宛先ポートに対応する要素に Δ を足す（ステップ 4）。

【 0 0 2 6 】

図 1 3 は、障害下にあったチャネル制御部が障害から回復した場合における S V P 4 0 が分配テーブル 3 1 を更新する処理の一例を示す流れ図である。S V P が更新する分配テーブル 3 1 の要素は転送先とする確率を示したものである場合の説明をする。S V P 4 0 の障害監視部 4 0 0 1 により、障害下にあったチャネル制御部が障害から回復したことを確認する（ステップ 1）。それを受けて、分配テーブル 3 1 の宛先が該当チャネル制御部の列について、ある定数 Δ 以上の要素に関しては Δ を引く。それに対応する障害回復チャネル制御部の列の要素について Δ を足す（ステップ 2）。

【 0 0 2 7 】

図 1 4 は、あるチャネル制御部 1 1 が障害に陥った場合に、該当チャネル制御部をフェイルオーバーする S V P 4 0 の処理の一例を示す流れ図である。S V P が更新する分配テーブル 3 1 の要素は転送先とする確率を示したものである場合の説明をする。ここでは障害下にあるチャネル制御部を障害チャネル制御部と呼ぶ。S V P 4 0 の障害監視部 4 0 0 1 により、障害チャネル制御部の存在を確認する（ステップ 1）。次に負荷監視部 4 0 0 2 により得られた負荷情報テーブル 4 0 1 から正常なチャネル制御部の負荷率を参照し、最低負荷率のチャネル制御部

を見つける（ステップ2）。その後、分配テーブル31を参照し要求転送先チャンネル制御部が障害チャンネル制御部の行を順次チェックする（ステップ3）。該当要素が0より大きい、すなわちチャンネル-DKC間スイッチ30が障害チャンネル制御部に要求を転送することになっていた場合は、該当要素を0として障害チャンネル制御部に要求を転送しないようにし、該当最低負荷率のチャンネル制御部の要素に、障害チャンネル制御部の元の要素を加算する（ステップ4）。ステップ3、ステップ4を該当行全体に対してチェック終了するまで続ける。

つぎに、代行引継ぎの制御方法を説明する。リードアクセス要求の宛先のチャンネル制御部11が高い負荷状態にあるために、チャンネル-DKC間スイッチ30により、他のチャンネル制御部にリード要求が転送され、リード要求を受信したチャンネル制御部11が処理を途中段階まで代行処理行い、その後本来の宛先チャンネル制御部が残りの段階の引継ぎ処理を行う場合について、図15にリード要求が転送されたチャンネル制御部の代行処理を、図16に本来の宛先チャンネル制御部が行う引継ぎ処理を流れ図で示した。ライトアクセス要求の場合に図15、図16に対応するのが図17、図18である。この方法で制御するチャンネル制御部を備える装置で、SVP40が、分配テーブル31を変更することによりチャンネル制御部の障害を回避しフェイルオーバを行うチャンネル制御部の代行レベルテーブル113を変更することで完全代行を指定する制御方法を流れ図で示したものが図19である。該当チャンネル制御部が障害から回復した場合のSVPの制御方法は完全代行による場合と同様で図10の流れ図で示される。また、チャンネル制御部間の負荷に偏りが生じた場合に、分配テーブル31を変更し高い負荷状態にあるチャンネル制御部に転送するアクセス要求の割合を減らし、かつ代行レベルテーブル113により負荷分散先チャンネル制御部の代行レベルを設定することにより負荷分散するSVPの制御方法を図21に示す。図19、図10、図21はいずれも分配テーブルの要素が転送先候補とするか否かの2値で示す場合の実施例であり、分配テーブルの要素を転送先とする確率を示す場合の対応する実施例は図22、図13、図23に示す。また、代行レベルを変更することにより、より細かい負荷分配の設定を行う場合のSVP40の制御方法を示す流れ図を図24に示す。

【 0 0 2 8 】

図 1 5 は、リード要求受信時の処理の一例を示す流れ図である。代行レベルテーブル 1 1 3 に設定された段階の処理までを要求受領チャンネル制御部が行い、以降の段階の処理を要求宛先チャンネル制御部が行うことが本実施例の特徴である。

【 0 0 2 9 】

リード要求を受領すると、受信要求から、本来の要求宛先のチャンネル制御部とコマンドとアクセス先アドレスを解析し、リードアクセスであることを認識する（ステップ 1）。アクセス先アドレスは、制御メモリ部 1 2 の装置構成情報 1 2 3 を参照することで、アクセス要求先のディスク制御装置番号とドライブ番号を識別できる。次に代行レベルテーブル 1 1 3 を参照し、該当チャンネル制御部に対する代行レベルを獲得する。該当代行レベルが 1 の場合は、要求宛先チャンネル制御部に代行処理情報 1 2 4 を示すことで処理の引継ぎを行い、要求受領チャンネル制御部は処理を終了する（ステップ 1 1）。本実施例では、制御メモリ部 1 2 を通じて代行処理情報 1 2 4 を示すが、メッセージ通信などで該当チャンネル制御部に直接代行処理情報 1 2 4 を示してもよい。該当代行レベルが 2 以上の場合は、ステップ 1 で識別した当該 D K C のキャッシュに対してキャッシュヒットミス判定を行う（ステップ 2）。キャッシュに保持していないキャッシュミスの場合は、当該 D K C のディスク制御部に対して当該データのドライブからキャッシュへの転送依頼を行う（ステップ 3）。ここで該当代行レベルが 2 の場合は、要求宛先チャンネル制御部に代行処理情報 1 2 4 を示すことで処理の引継ぎを行い、要求受領チャンネル制御部は処理を終了する（ステップ 1 3）。該当代行レベルが 3 以上の場合は、転送終了までリード処理を中断し（ステップ 6）、ステージング終了後、再びリード処理を継続することになる。ここで該当代行レベルが 3 の場合は、要求宛先チャンネル制御部に代行処理情報 1 2 4 を示すことで処理の引継ぎを行い、要求受領チャンネル制御部は処理を終了する（ステップ 1 4）。ステップ 3 でヒット判定の場合、該当代行レベルが 2 以上の場合は、要求宛先チャンネル制御部に代行処理情報 1 2 4 を示すことで処理の引継ぎを行い、要求受領チャンネル制御部は処理を終了する（ステップ 1 5）。その後、または、ステップ 4 でステージング処理が終了した場合は、ホストコンピュータに対して当該データを転送する

(ステップ5)。ステップ5の際、処理を行ったチャネル制御部がホストコンピュータのアクセス要求の宛先と異なっているとしても、本来の宛先チャネル制御部が応答したとるようにデータを転送する。

【0030】

図16は、図15に対応するリード要求に対する代行処理の引継ぎ処理の一例を示す流れ図である。リード要求を受領した代行チャネル制御部の代行レベルテーブル113に設定された段階の処理までを代行チャネル制御部が行い、以降の段階の処理を要求宛先チャネル制御部が行うこと（引継ぎ処理）が本実施例の特徴である。

【0031】

チャネル制御部が該当チャネル制御部を対象とした代行処理情報124の存在を確認すると、該当代行処理情報を解析する（ステップ1）。代行処理情報を解析し処理段階1245を取り出すことにより、代行レベルが判別できる。代行レベルが1ならば、アクセス先DKCのキャッシュに対してキャッシュヒット／ミス判定を行う（ステップ2）。代行レベルが2以上ならば、処理段階1245よりキャッシュヒット／ミスが判別できる。ステップ2でキャッシュミスもしくは代行レベル2でキャッシュミスの場合は、該当ディスク制御部に該当データをキャッシュに転送依頼する（ステップ3）。代行レベル3でキャッシュミスまたはステップ3終了後は、転送終了までリード処理を中断し（ステップ4）、ステージング終了後、再びリード処理を継続することになる。キャッシュヒットもしくはステップ4終了後、該当DKCのチャネル制御部に対し、該当データを参照しチャネルに転送する（ステップ5）。

【0032】

図17は、ライト要求受信時の処理の一例を示す流れ図である。代行レベルテーブル113に設定された段階の処理までを要求受領チャネル制御部が行い、以降の段階の処理を要求宛先チャネル制御部が行うことが本実施例の特徴である。

【0033】

ライト要求を受領すると、受信要求から、本来の要求宛先のチャネル制御部とコマンドとアクセス先アドレスを解析し、ライトアクセスであることを認識する

(ステップ1)。アクセス先アドレスは、制御メモリ部12の装置構成情報123を参照することで、アクセス要求先のディスク制御装置番号とドライブ番号を識別できる。次に代行レベルテーブル113を参照し、該当チャネル制御部に対する代行レベルを獲得する。該当代行レベルが1の場合は、要求宛先チャネル制御部に代行処理情報124を示すことで処理の引継ぎを行い、要求受領チャネル制御部は処理を終了する(ステップ11)。本実施例では、制御メモリ部12を通じて代行処理情報124を示すが、メッセージ通信などで該当チャネル制御部に直接代行処理情報124を示してもよい。該当代行レベルが2以上の場合は、ステップ1で識別した当該DKCのキャッシュに対してキャッシュヒットミス判定を行う(ステップ2)。キャッシュに保持していないキャッシュミスの場合は、当該DKCのディスク制御部に対して当該データのドライブからキャッシュへの転送依頼を行う(ステップ3)。ここで該当代行レベルが2の場合は、要求宛先チャネル制御部に代行処理情報124を示すことで処理の引継ぎを行い、要求受領チャネル制御部は処理を終了する(ステップ13)。該当代行レベルが3以上の場合は、転送終了までライト処理を中断し(ステップ6)、ステージング終了後、再びライト処理を継続することになる。ここで該当代行レベルが3の場合は、要求宛先チャネル制御部に代行処理情報124を示すことで処理の引継ぎを行い、要求受領チャネル制御部は処理を終了する(ステップ14)。ステップ3でヒット判定の場合、該当代行レベルが2または3の場合は、要求宛先チャネル制御部に代行処理情報124を示すことで処理の引継ぎを行い、要求受領チャネル制御部は処理を終了する(ステップ15)。その後、または、ステップ4でステージング処理が終了した場合は、当該DKCのキャッシュに対して当該データの更新を行う(ステップ5)。該当代行レベルが4の場合は、要求宛先チャネル制御部に代行処理情報124を示すことで処理の引継ぎを行い、要求受領チャネル制御部は処理を終了する(ステップ16)。更新終了後、ホストコンピュータに対してライト処理の完了報告を行う(ステップ6)。ステップ6の際、処理を行ったチャネル制御部がホストコンピュータのアクセス要求の宛先と異なっている場合、本来の宛先チャネル制御部が応答したとする完了報告をする。

【0034】

図 1 8 は、図 1 7 に対応するライト要求に対する代行処理の引継ぎ処理の一例を示す流れ図である。ライト要求を受領した代行チャネル制御部の代行レベルテーブル 1 1 3 に設定された段階の処理までを代行チャネル制御部が行い、以降の段階の処理を要求宛先チャネル制御部が行うこと（引継ぎ処理）が本実施例の特徴である。

【 0 0 3 5 】

チャネル制御部が該当チャネル制御部を対象とした代行処理情報 1 2 4 の存在を確認すると、該当代行処理情報を解析する（ステップ 1）。代行処理情報を解析し処理段階 1 2 4 5 を取り出すことにより、代行レベルが判別できる。代行レベルが 1 ならば、アクセス先 DKC のキャッシュに対してキャッシュヒット／ミス判定を行う（ステップ 2）。代行レベルが 2 以上ならば、処理段階 1 2 4 5 よりキャッシュヒット／ミスが判別できる。ステップ 2 でキャッシュミスもしくは代行レベル 2 でキャッシュミスの場合は、該当ディスク制御部に該当データをキャッシュに転送依頼する（ステップ 3）。代行レベル 3 でキャッシュミスまたはステップ 3 終了後は、転送終了までライト処理を中断し（ステップ 4）、ステージング終了後、再びライト処理を継続することになる。代行レベル 3 でキャッシュヒットもしくはステップ 4 終了後、該当 DKC のキャッシュに対し該当データを更新する（ステップ 5）。ステップ 5 終了後もしくは代行レベル 4 の場合はホストにライト完了報告する（ステップ 6）。

【 0 0 3 6 】

図 1 9 は、あるチャネル制御部 1 1 が障害に陥った場合に、該当チャネル制御部をフェイルオーバーする SVP 4 0 の処理の一例を示す流れ図である。SVP が更新する分配テーブル 3 1 の要素は転送先候補とするか否かの 2 値であり、更に各チャネル制御部が代行レベルテーブル 1 1 3 を持ち SVP により更新される場合の説明をする。ここでは障害下にあるチャネル制御部を障害チャネル制御部と呼ぶ。SVP 4 0 の障害監視部 4 0 0 1 により、障害チャネル制御部の存在を確認する（ステップ 1）。次に負荷監視部 4 0 0 2 により得られた正常なチャネル制御部の負荷率を障害情報テーブル 4 0 1 より参照し、最低負荷率のチャネル制御部を見つける（ステップ 2）。その後、分配テーブル 3 1 を参照し要求転送先

チャンネル制御部が障害チャンネル制御部の行を順次チェックする（ステップ3）。該当要素が1、すなわちチャンネル-DKC間スイッチ30が障害チャンネル制御部に要求を転送することになっていた場合は、該当要素を0として障害チャンネル制御部に要求を転送しないようにし、また、該当要素の列の要求転送先チャンネル制御部が該当最低負荷率のチャンネル制御部の要素を1にして要求転送先チャンネル制御部が無い状態を防ぐ（ステップ4）。また、該当最低負荷率のチャンネル制御部が有する代行レベルテーブル113の障害チャンネル制御部に対応する要素を完全代行とする（ステップ5）。ステップ3、ステップ4、ステップ5を該当行全体に対してチェック終了するまで続ける。

【0037】

図21は、各チャンネル制御部11の負荷に偏りが生じた場合におけるSVP40の処理の一例を示す流れ図である。SVP40の負荷監視部4002により得られた負荷情報テーブル401から各チャンネル制御部の負荷率に偏りがあることを確認する（ステップ1）。負荷率の偏りは、例えば、最高負荷率と最低負荷率の差が閾値を超えた場合で定義できる。次に、チャンネル制御部を負荷率昇順でソートする（ステップ2）。その結果により最高負荷率のチャンネル制御部の負荷を下げるように分配テーブル31を更新する。分配テーブルで要求宛先チャンネル制御部が該当最高負荷率のチャンネル制御部の列をステップ2のソート順にチェックしていく（ステップ3）。該当要素が0ならば、該当転送先チャンネル制御部が有する代行レベルテーブル113の最高負荷率チャンネル制御部の代行レベルを代行時の初期値に設定し（ステップ4）、該当要素を1として負荷分散先として登録する（ステップ5）。該当要素が1ならば次の要素をチェックする。全要素チェック後またはステップ5終了後は、分配テーブルで要求転送先チャンネル制御部が該当最高負荷率のチャンネル制御部である行をステップ2のソート順にチェックしていく（ステップ6）。該当要素が1ならば、該当要素を0として該当最高負荷率のチャンネル制御部を負荷分散先から削除し（ステップ7）、最高負荷率チャンネル制御部の代行レベルテーブルの該当宛先チャンネル制御部の代行レベルを代行しないに設定する（ステップ8）。該当要素が0ならば次の要素をチェックする。全要素チェック後またはステップ8終了後は、この処理を終了とする。

【 0 0 3 8 】

図 2 2 は、あるチャネル制御部 1 1 が障害に陥った場合に、該当チャネル制御部をフェイルオーバーする S V P 4 0 の処理の一例を示す流れ図である。S V P が更新する分配テーブル 3 1 の要素は転送先とする確率を示したものであり、更に各チャネル制御部が代行レベルテーブル 1 1 3 を持ち S V P により更新される場合の説明をする。ここでは障害下にあるチャネル制御部を障害チャネル制御部と呼ぶ。S V P 4 0 の障害監視部 4 0 0 . 1 により、障害チャネル制御部の存在を確認する（ステップ 1）。次に負荷監視部 4 0 0 2 により得られた正常なチャネル制御部の負荷率を障害情報テーブル 4 0 1 より参照し、最低負荷率のチャネル制御部を見つける（ステップ 2）。その後、分配テーブル 3 1 を参照し要求転送先チャネル制御部が障害チャネル制御部の行を順次チェックする（ステップ 3）。該当要素が 0 より大きい、すなわちチャネル-D K C 間スイッチ 3 0 が障害チャネル制御部に要求を転送することになっていた場合は、該当要素を 0 として障害チャネル制御部に要求を転送しないようにし、該当最低負荷率のチャネル制御部の要素に、障害チャネル制御部の元の要素を加算する（ステップ 4）。また、該当最低負荷率のチャネル制御部が有する代行レベルテーブル 1 1 3 の障害チャネル制御部に対応する要素を完全代行とする（ステップ 5）。ステップ 3、ステップ 4、ステップ 5 を該当行全体に対してチェック終了するまで続ける。

【 0 0 3 9 】

図 2 3 は、各チャネル制御部 1 1 の負荷に偏りが生じた場合における S V P 4 0 の処理の一例を示す流れ図である。S V P が更新する分配テーブル 3 1 の要素は転送先とする確率を示したものであり、更に各チャネル制御部が代行レベルテーブル 1 1 3 を持ち S V P により更新される場合の説明をする。S V P 4 0 の負荷監視部 4 0 0 2 により得られた負荷情報テーブル 4 0 1 から各チャネル制御部の負荷率に偏りがあることを確認する（ステップ 1）。次に正常チャネル制御部の負荷率が最低であるものと最高であるものを選択する（ステップ 2）。分配テーブル 3 1 中の最高負荷率のチャネル制御部が転送先ポートの行について全ての Δ 以上の要素から Δ を減算する（ステップ 3）。この際、障害下にあるチャネル制御部に該当する要素については除外する。ここで、代行レベルテーブル 1 1 3

についての整合を取る。ステップ 3 にて 0 になった要素に対応する最高負荷率チャンネル制御部の代行レベルテーブルの該当宛先チャンネル制御部の代行レベルを代行しないに設定する（ステップ 4）。次に、分配テーブル 3 1 中の最低負荷率のチャンネル制御部が転送先ポートの行についてステップ 3 にて減算した宛先ポートに対応する要素に Δ を足す（ステップ 5）。またここでも、代行レベルテーブル 1 1 3 についての整合を取る。ステップ 5 にて 0 から増加した要素に対応する転送先チャンネル制御部の代行レベルテーブルについて最高負荷率チャンネル制御部の代行レベルを代行時の初期値に設定する（ステップ 6）。

【 0 0 4 0 】

図 2 4 は、各チャンネル制御部 1 1 の負荷に偏りが生じた場合における S V P 4 0 が各チャンネル制御部の代行レベルテーブル 1 1 3 の変更を行う処理の一例を示す流れ図である。S V P 4 0 の負荷監視部 4 0 0 2 により得られた負荷情報テーブル 4 0 1 から各チャンネル制御部の負荷率に偏りがあることを確認する（ステップ 1）。次に正常チャンネル制御部の負荷率が最低であるものと最高であるものを選択する（ステップ 2）。ここで、分配テーブル 3 1 を参照し、最低負荷率のチャンネル制御部が最高負荷率のチャンネル制御部を代行しているように記載されているか調べる（ステップ 3）。まだ記載が無ければ、前述の実施例により分配テーブル 3 1 を更新し処理を終了する（ステップ 6）。すでに有るならば、最低負荷率のチャンネル制御部の代行レベルテーブル 1 1 3 を参照し、現在の最高負荷率チャンネル制御部に対する代行レベルを調査する（ステップ 4）。既に最高レベルならばこの処理は終了する。まだ最高レベルに至っていない場合には、最低負荷率チャンネル制御部の最高負荷率チャンネル制御部に対する代行レベルを 1 段階上げる（ステップ 5）。

【 0 0 4 1 】

以上で、チャンネル制御部 1 1 および S V P 4 0 の制御方法を説明した。次に分配テーブルの要素が転送先確率を示す場合におけるチャンネル - D K C 間スイッチ 3 0 のアクセス要求の転送制御方法を図 2 5 により説明する。

【 0 0 4 2 】

図 2 5 は、分配テーブル 3 1 の要素が転送先とする確率を示す場合のチャンネル

—DKC間スイッチ30において、ホスト側ポート3011がアクセス要求を受けた場合どのDKC側ポート3021に転送するかを選択方法を流れ図で示したものである。まず、ホストコンピュータ0のアクセス要求がホスト側ポート3011に到着したことをホスト側ポートが確認する（ステップ1）。次に、該当アクセス要求の宛先ポートを解析する（ステップ2）。更に、擬似乱数を発生する（ステップ3）。擬似乱数は、つねにカウントアップしているカウンタ値や、毎回更新される乱数の種に一定規則の演算を施し、その結果の下位を用いるなどの方法がある。次に、分配テーブル31の該当宛先ポートの列を参照し、要素がステップ3で得た擬似乱数以上のものを転送先候補とする（ステップ4）。もし候補数が0ならばステップ3に戻る（ステップ5）。候補があるならば、候補の中で、前回選択されたもののインデックスから周期境界条件で正の方向に最も近いものを転送先ポート番号として選択し、該当要求を選択したDKC側ポートに転送する（ステップ6）。

【0043】

次に、装置管理者がSVP40を通じて装置の負荷分散に関する設定を行う場合の画面の一例について図26により説明する。

【0044】

図26は、SVP40がチャネル制御部に対して指示を行うことにより実行する負荷分散について、SVP40の装置管理インタフェース部404を通じて端末（図示せず）等で装置管理者が設定を行う場合の設定画面の実施例である。代行レベル設定画面2601は設定するチャネル制御部を選択する設定チャネル制御部選択欄2630と各チャネル制御部に対応した代行レベル選択欄2610を有する。装置管理者は、キーボード入力やGUIによる選択入力等の方法で、設定チャネル制御部選択欄2630や代行レベル選択欄2610にて、設定するチャネル制御部を選択し、各チャネル制御部の代行レベルを設定可能である。これらは、設定ボタン2621を前述した入力方法で選択すると、本画面で設定した内容が、SVP40を通じて代行レベルテーブル113と分配テーブル31に反映される。代行レベル選択欄2610にて自動設定に選択されたチャネル制御部は、SVP40が代行レベルの更新を行う。これら設定された情報はSVP40

のローカルメモリ 4 0 0 4 に収納され次回に本画面にて再設定されるまで保持される。また、取消ボタン 2 6 2 2 を前述した入力方法で選択すると、本画面を呼び出す前の設定情報が引き続き使用される。

【 0 0 4 5 】

【発明の効果】

チャンネル-DKC間スイッチを設け、アクセス要求を複数のチャンネル制御部に振り分けることを可能にした。このことにより、ホストコンピュータ側に装置の内部構造を意識させないで、チャンネル制御部の負荷に応じた負荷分散が可能になる。同様に、障害下にあるチャンネル制御部を迂回してアクセス要求を他のチャンネル制御部で処理することができる。この場合についても、ホストコンピュータ側は該当チャンネル制御部が障害に陥ったことや装置内部構造を意識せずに操作を継続することができる。

【 0 0 4 6 】

さらに、本発明の代行処理により負荷分散を行えば、設定した代行レベルによっては、リードアクセス要求処理に、処理を引継ぐチャンネル制御部のDKCに対象データが存在するとすれば、リードデータがディスク制御装置間の通信手段を経ることが無くホストコンピュータに応答することが可能になる。これにより、ディスク制御装置間の通信手段に対する負荷分散にもなる。特にシーケンシャルリードの要求が多い環境において負荷分散をする場合、完全代行のような方法ではリードデータがディスク制御装置間の通信手段を占有し、ディスク制御装置間の通信手段がボトルネックとなってしまう。しかしながら、代行処理を用いることにより、装置使用環境に特徴的なアクセス種類によらず、チャンネル制御部の負荷分散が可能になる。

【図面の簡単な説明】

【図 1】

本発明に係わるディスク制御装置の概要を示すブロック図の一例である。

【図 2】

従来のディスク制御装置の概要を示すブロック図の一例である。

【図 3】

従来のディスク制御装置の概要を示すブロック図の一例である。

【図 4】

本発明のディスク制御装置のチャンネル-DKC間スイッチを示すブロック図の一例である。

【図 5】

本発明に係わるディスク制御装置のチャンネル制御部を示すブロック図の一例である。

【図 6】

本発明に係わるディスク制御装置の制御メモリを示すブロック図の一例である。

【図 7】

本発明に係わるディスク制御装置のチャンネル制御部の完全代行リード要求処理の一例を示す流れ図である。

【図 8】

本発明に係わるディスク制御装置のチャンネル制御部の完全代行ライト要求処理の一例を示す流れ図である。

【図 9】

本発明に係わるディスク制御装置のSVPがチャンネル制御部障害によるフェイルオーバーの処理の一例を示す流れ図である。

【図 1 0】

本発明に係わるディスク制御装置のSVPによるチャンネル制御部障害回復した場合の処理の一例を示す流れ図である。

【図 1 1】

本発明に係わるディスク制御装置のSVPがチャンネル制御部の負荷分散を行う場合の処理の一例を示す流れ図である。

【図 1 2】

本発明に係わるディスク制御装置のSVPがチャンネル制御部の負荷分散を行う場合の処理の一例を示す流れ図である。

【図 1 3】

本発明に係わるディスク制御装置のSVPによるチャネル制御部障害回復した場合の処理の一例を示す流れ図である。

【図 1 4】

本発明に係わるディスク制御装置のSVPによるチャネル制御部障害のフェイルオーバーの処理の一例を示す流れ図である。

【図 1 5】

本発明に係わるディスク制御装置のチャネル制御部の代行リード要求処理の一例を示す流れ図である。

【図 1 6】

本発明に係わるディスク制御装置のチャネル制御部の代行引継ぎリード要求処理の一例を示す流れ図である。

【図 1 7】

本発明に係わるディスク制御装置のチャネル制御部の代行ライト要求処理の一例を示す流れ図である。

【図 1 8】

本発明に係わるディスク制御装置のチャネル制御部の代行引継ぎライト要求処理の一例を示す流れ図である。

【図 1 9】

本発明に係わるディスク制御装置のSVPによるチャネル制御部障害のフェイルオーバーの処理の一例を示す流れ図である。

【図 2 0】

本発明に係わるディスク制御装置のSVPを示すブロック図の一例である。

【図 2 1】

本発明に係わるディスク制御装置のSVPがチャネル制御部の負荷分散を行う場合の処理の一例を示す流れ図である。

【図 2 2】

本発明に係わるディスク制御装置のSVPによるチャネル制御部障害のフェイルオーバーの処理の一例を示す流れ図である。

【図 2 3】

本発明に係わるディスク制御装置のSVPがチャンネル制御部の負荷分散を行う場合の処理の一例を示す流れ図である。

【図 2 4】

本発明に係わるディスク制御装置のSVPがチャンネル制御部の負荷分散を行う場合の処理の一例を示す流れ図である。

【図 2 5】

本発明に係わるディスク制御装置のチャンネル-DKC間スイッチのアクセス要求転送先ポートを決定する処理の一例を示す流れ図である。

【図 2 6】

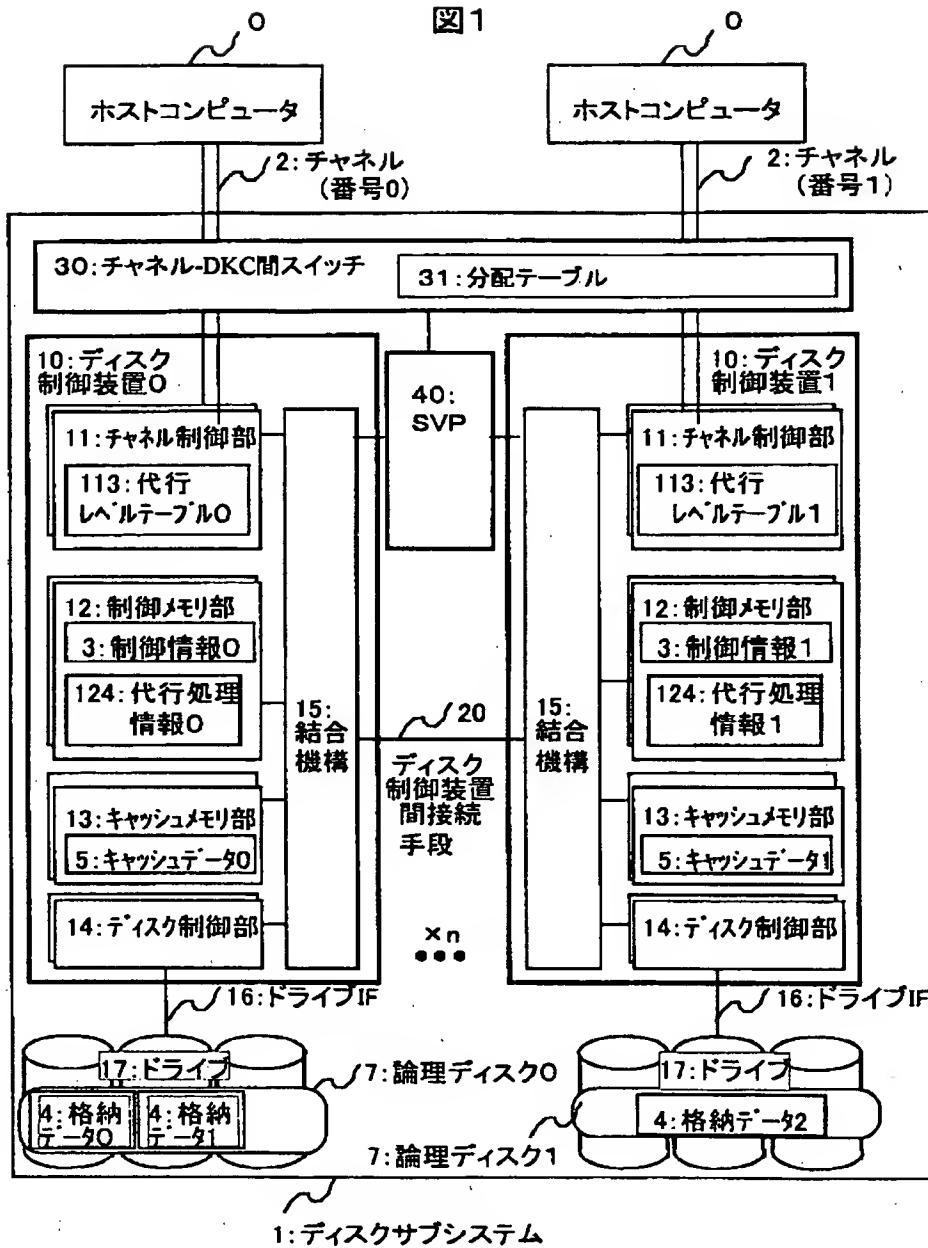
本発明に係わるディスク制御装置の代行レベルを設定時にSVPにより表示される画面の一例を示すブロック図である。

【符号の説明】

0・・・ホストコンピュータ、1・・・ディスクサブシステム、2・・・チャンネル、3・・・制御情報、4・・・格納データ、5・・・キャッシュデータ、7・・・論理ディスク、10・・・ディスク制御装置、11・・・チャンネル制御部、12・・・制御メモリ、13・・・キャッシュメモリ部、14・・・ディスク制御部、15・・・結合機構、16・・・ドライブIF、17・・・ドライブ、20・・・ディスク制御装置間接続手段、30・・・チャンネル-DKC間スイッチ、31・・・分配テーブル、39・・・SANスイッチ、40・・・SVP。

【書類名】 図面

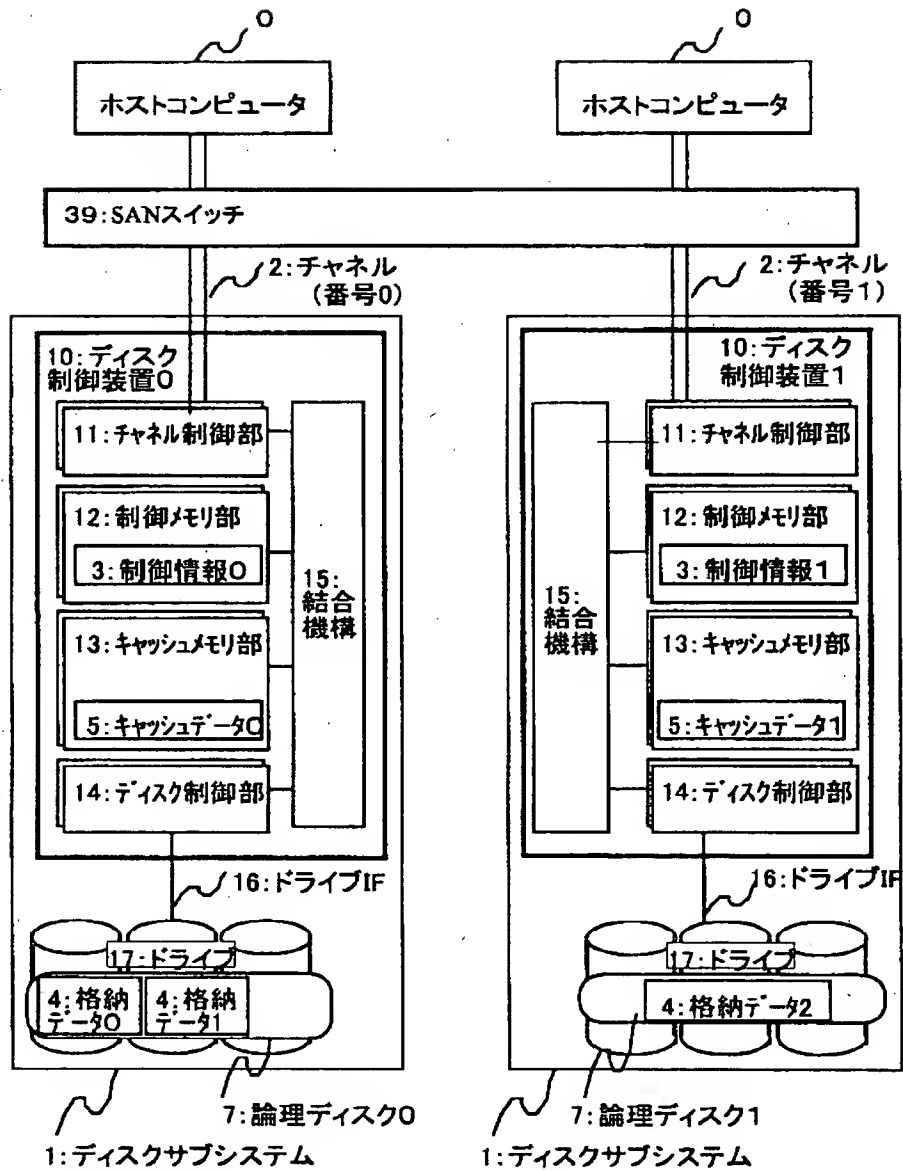
【図 1】



【図 2】

図 2

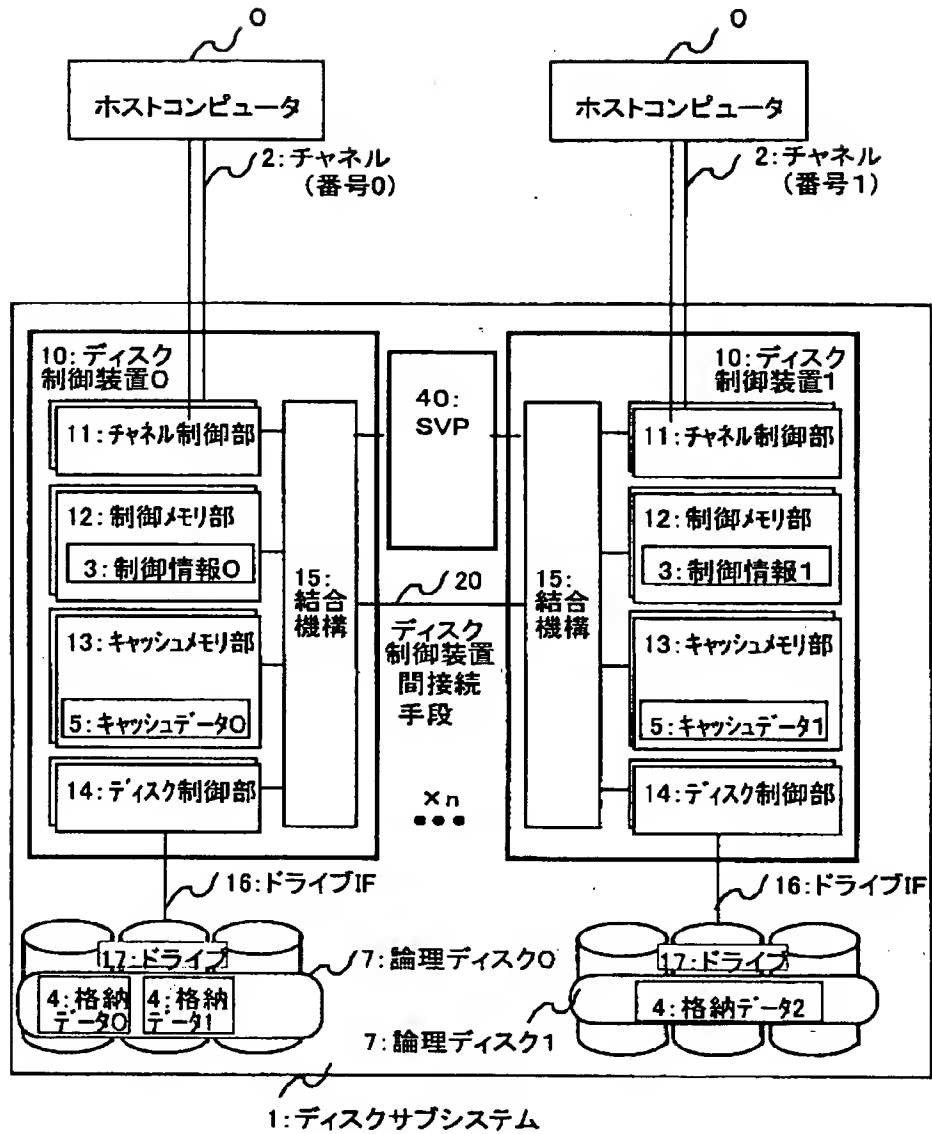
従来例



【図3】

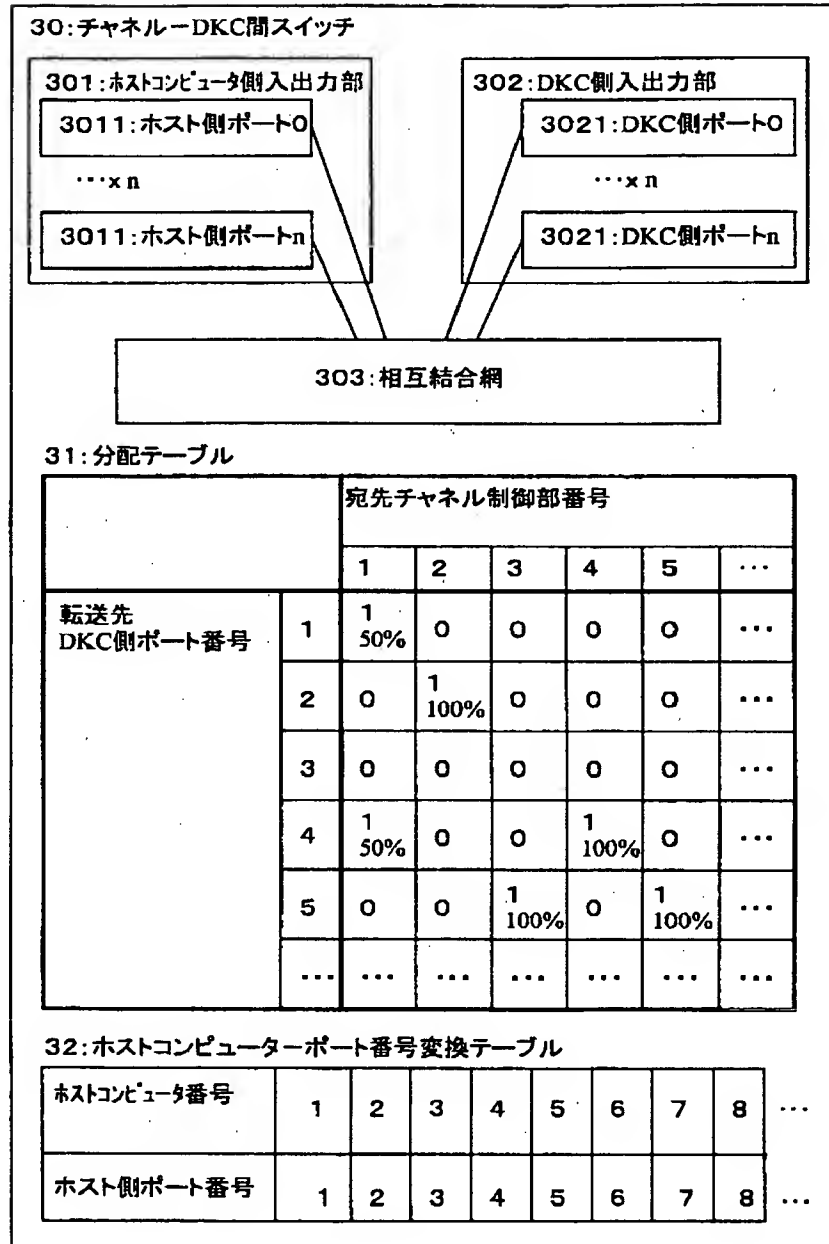
図3

従来例



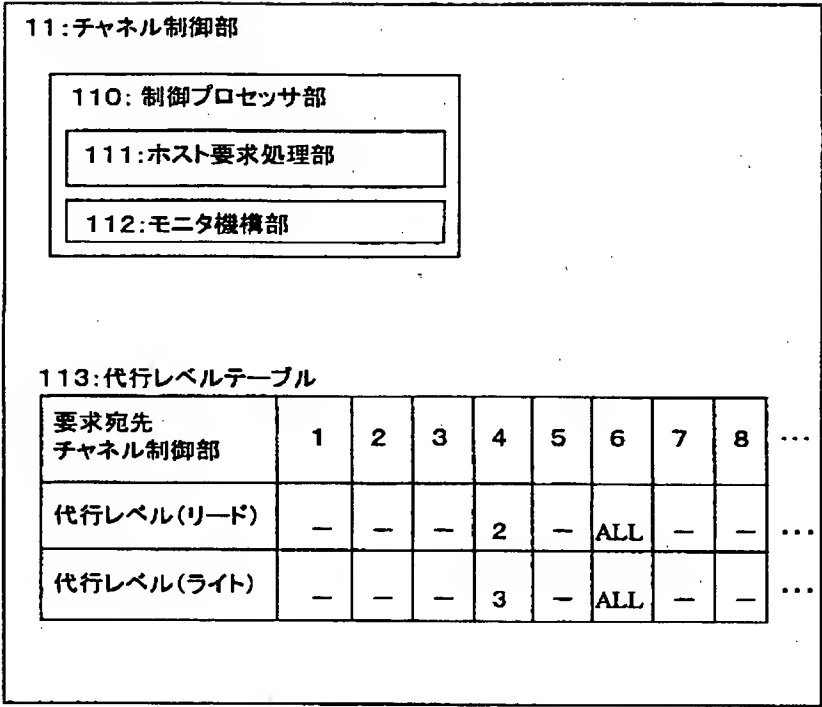
【図 4】

図4



【図 5】

図 5



【図 6】

図6

12:制御メモリ

121:制御情報

122:ディレクトリ情報

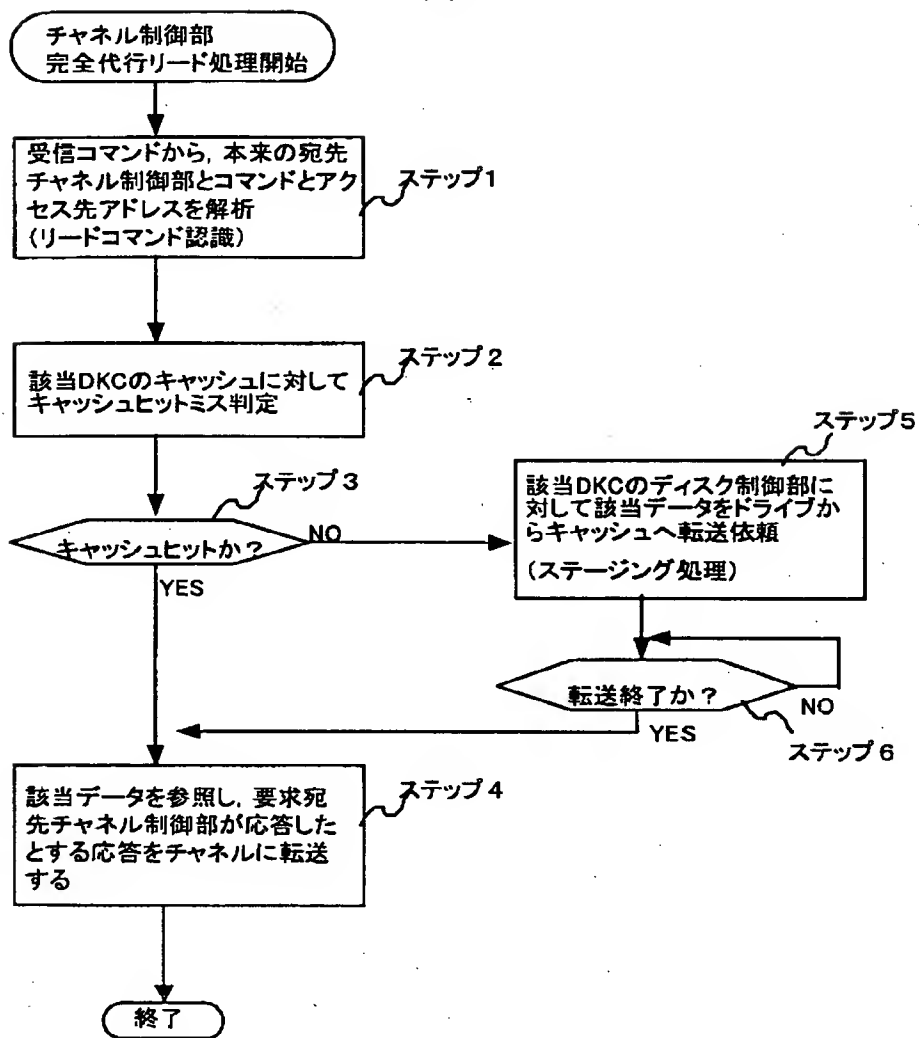
123:装置構成情報

124:代行処理情報

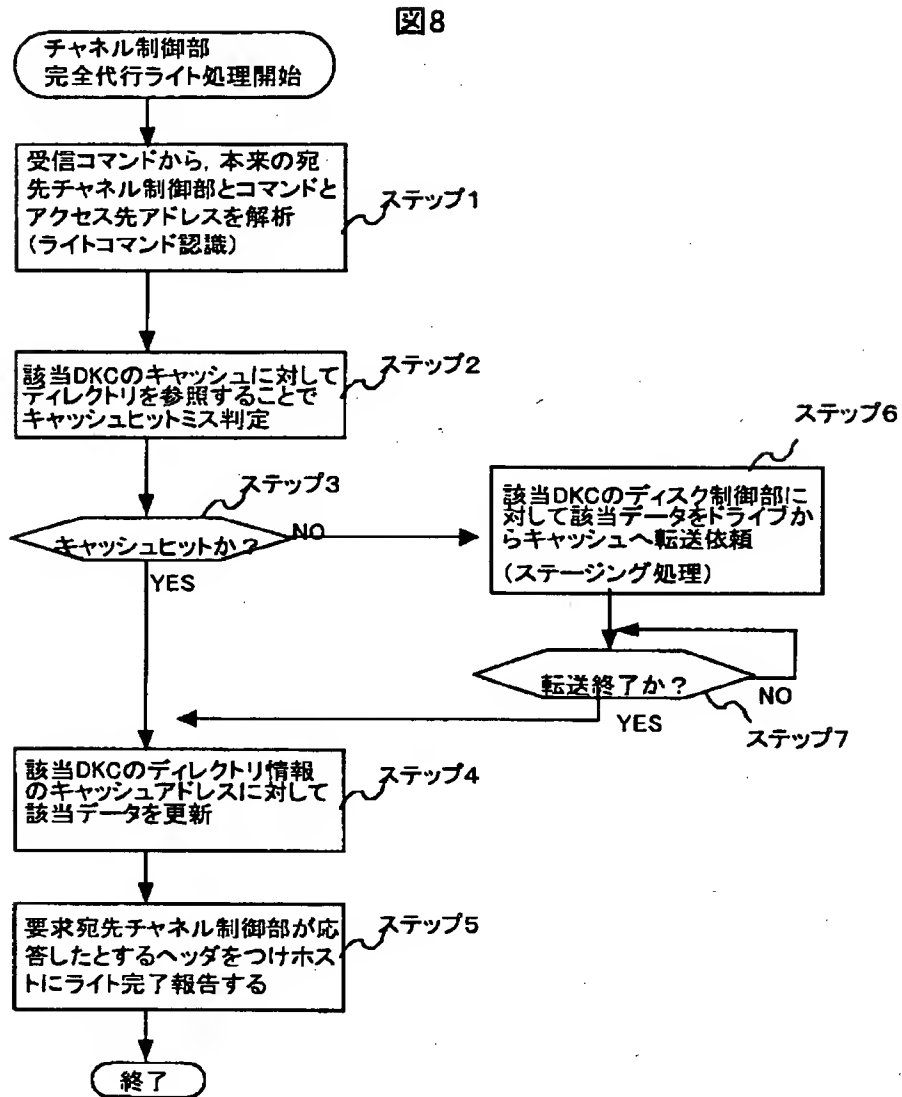
1241: ジョブ番号	1242: 要求ホスト	1243: 要求宛先	1244: コマンド内容	1245: 処理段階	1246: データ存在アドレス
101	ホスト0	チャネル 制御部3	リード	段階3 キャッシュミス	18000000H

【図 7】

図 7

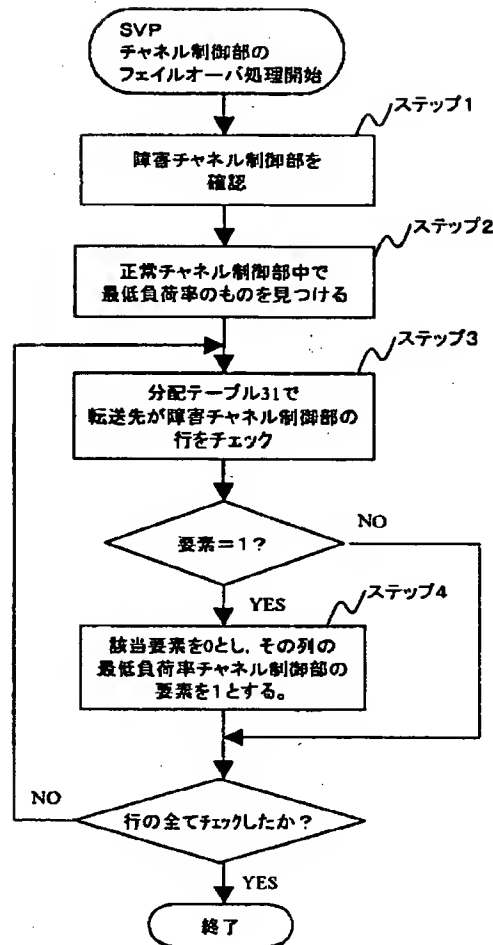


【図 8】



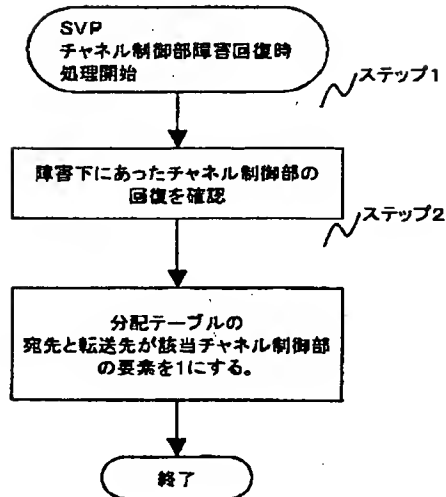
【図9】

図9



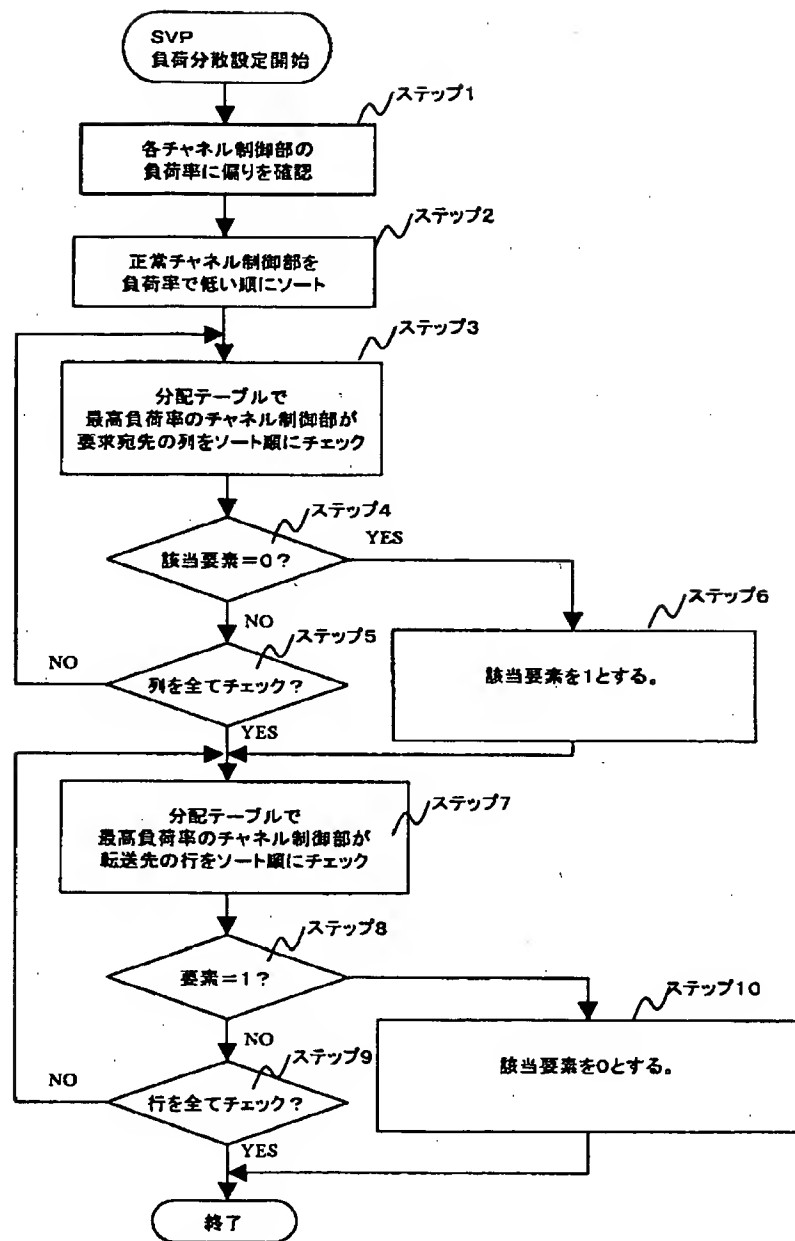
【図10】

図10



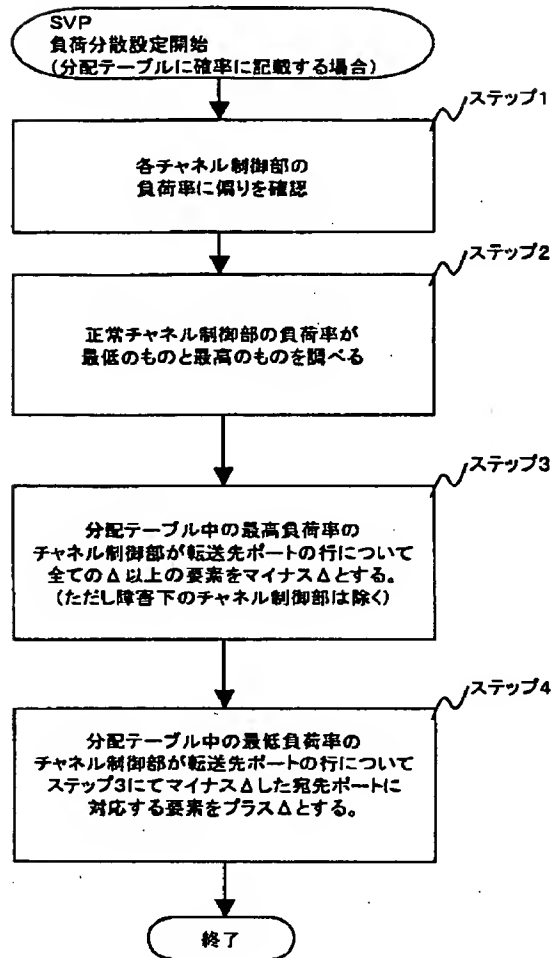
【図11】

図11



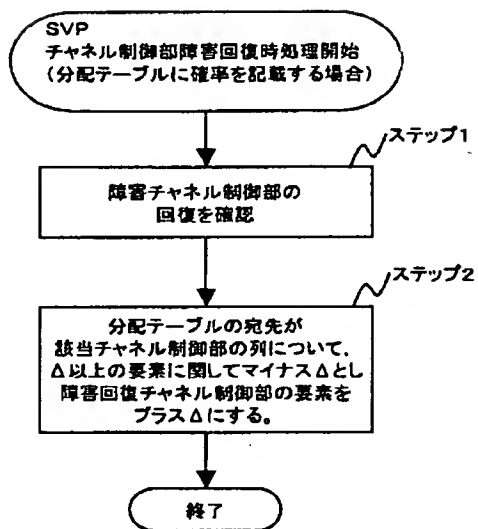
【図 12】

図12



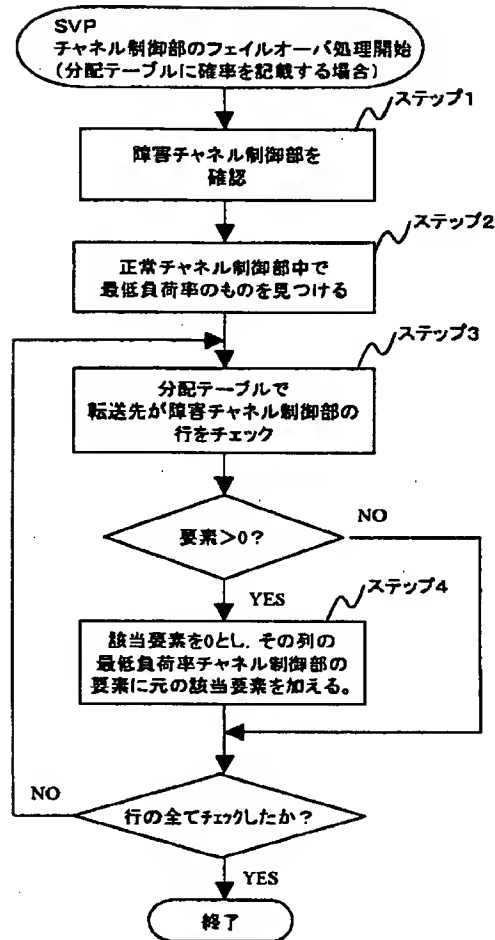
【図 13】

図 13



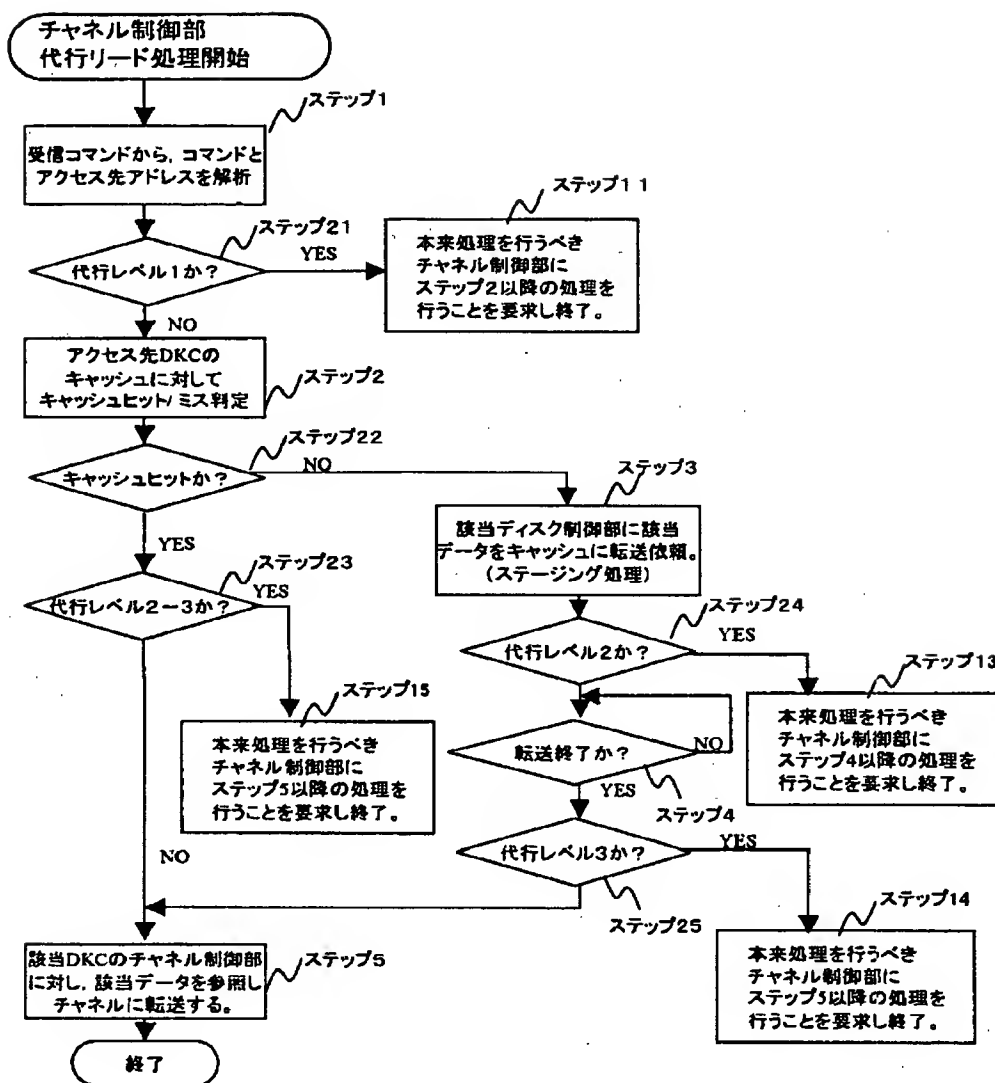
【図 14】

図14



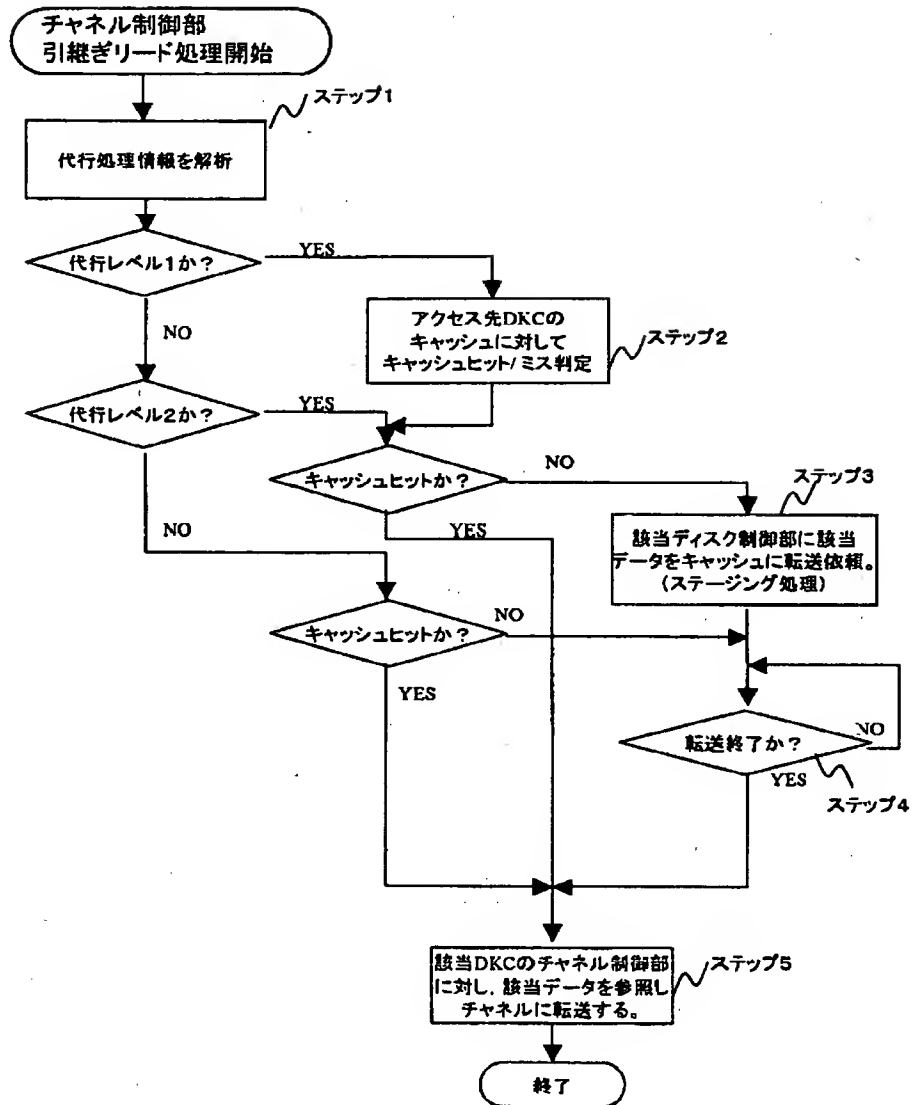
【図15】

図15

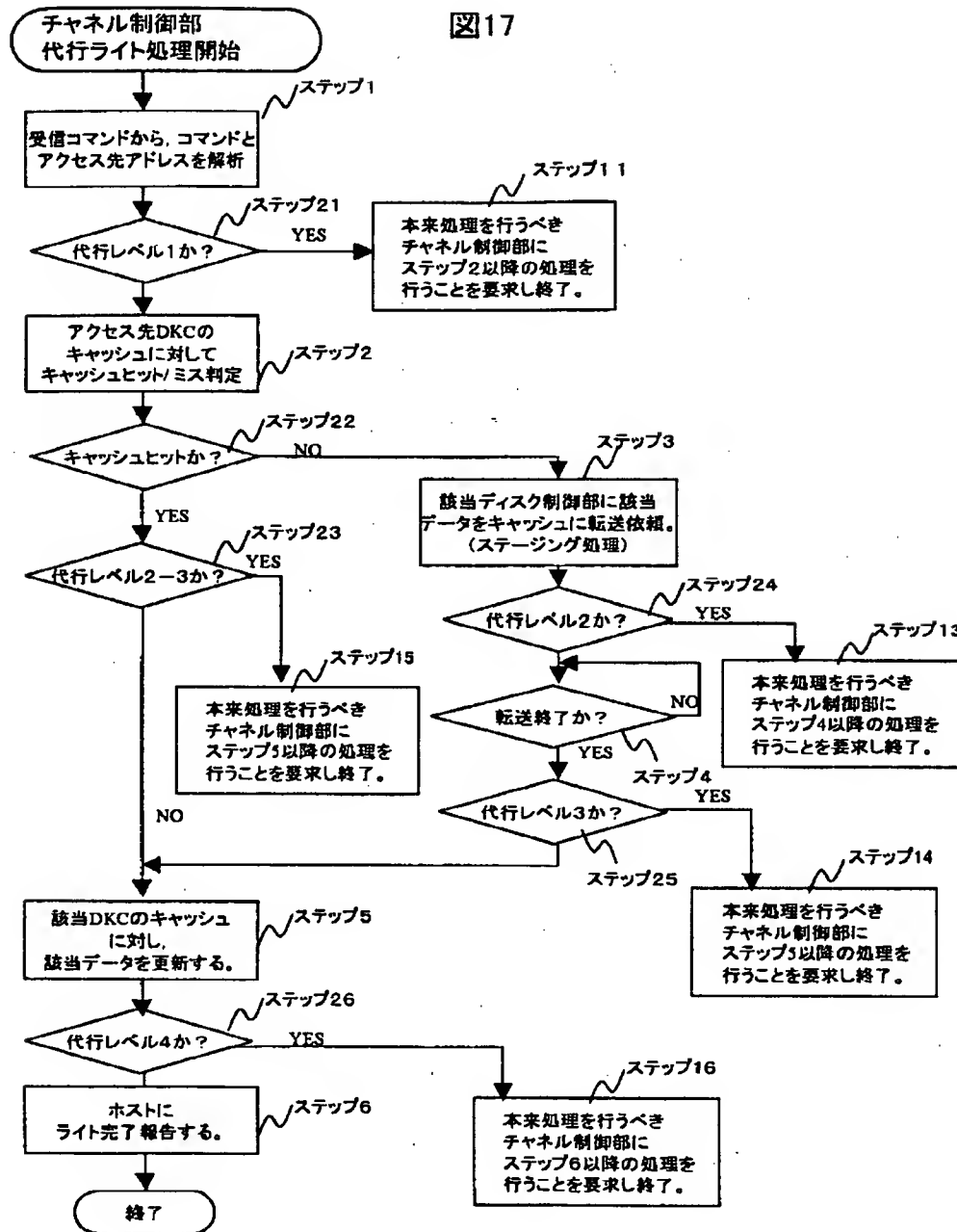


【図16】

図16

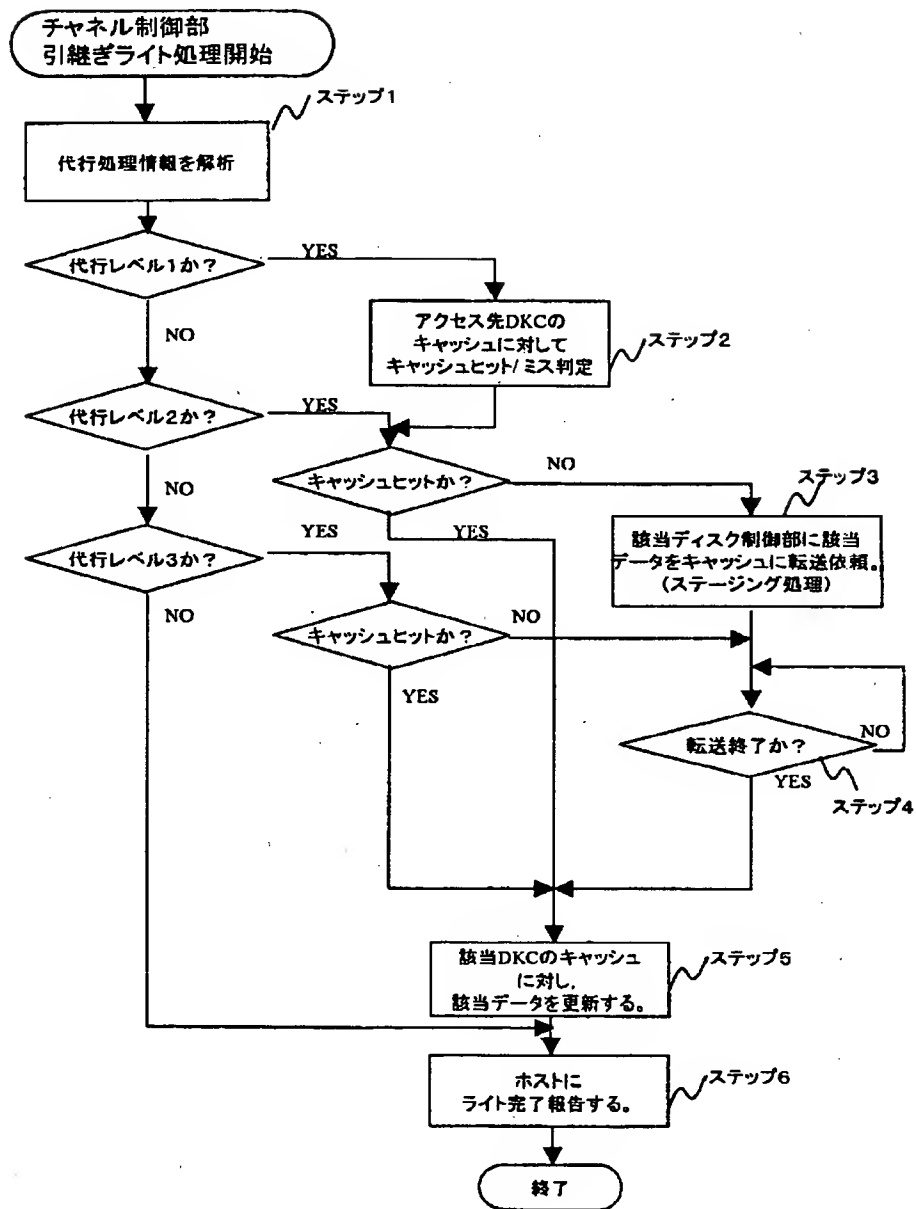


【図 17】



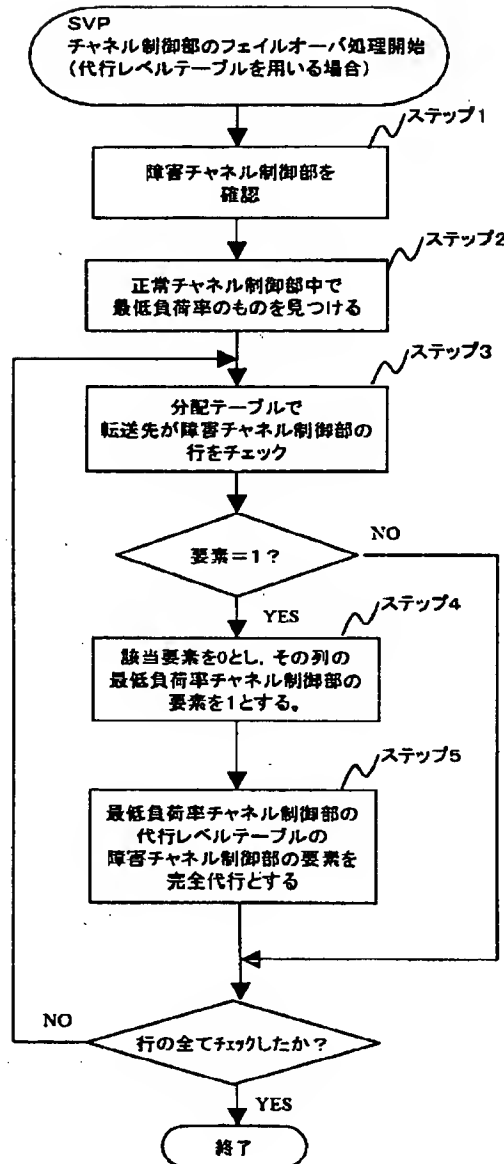
【図18】

図18



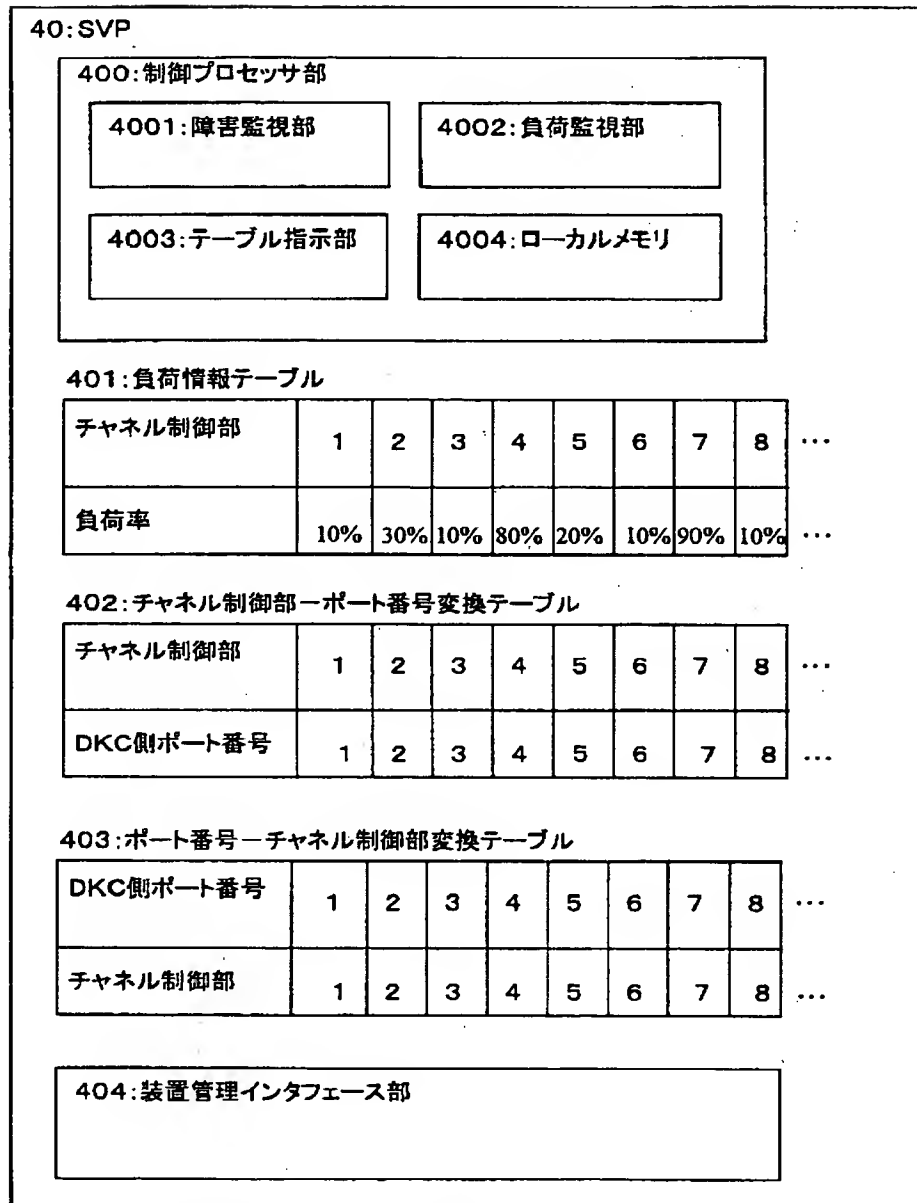
【図 19】

図 19



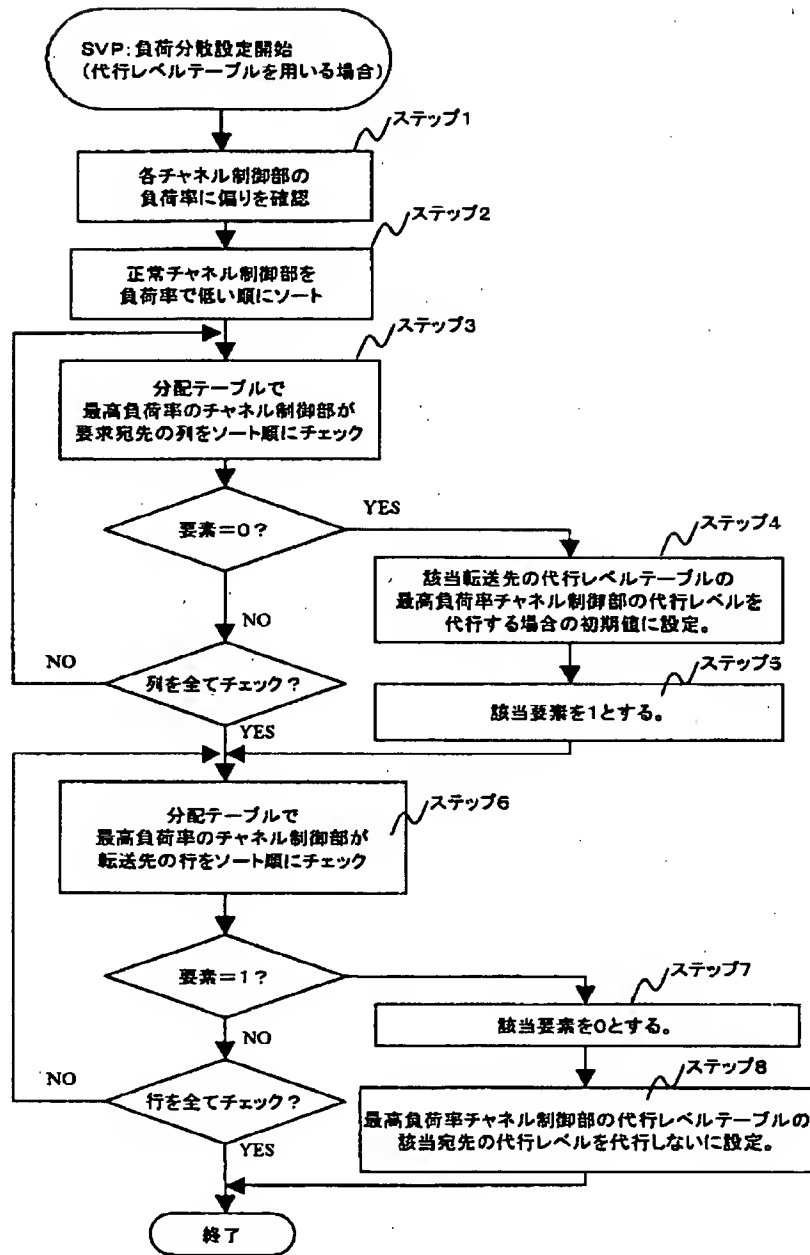
【図 2 0】

図20



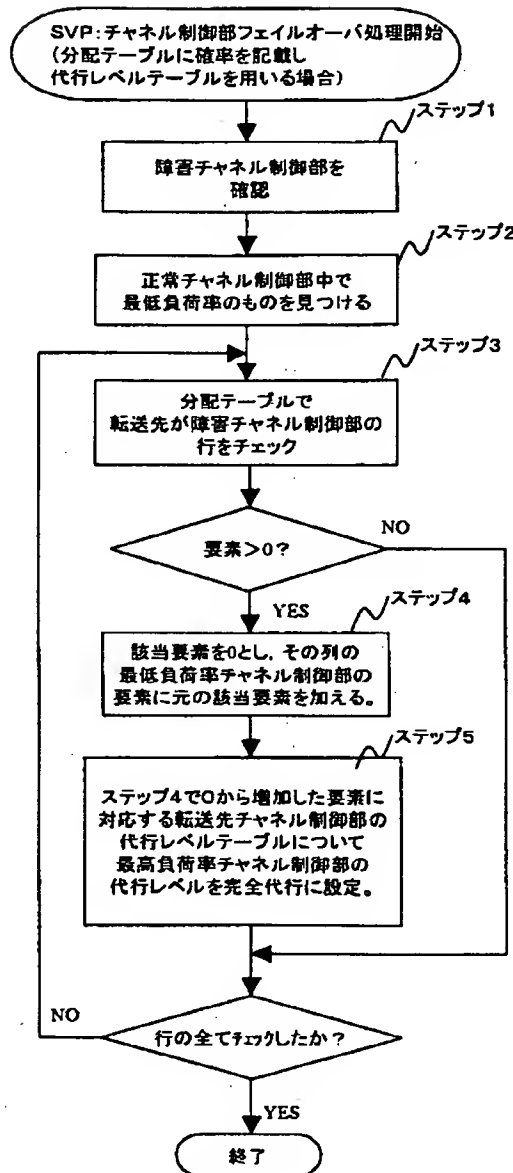
【図 21】

図21



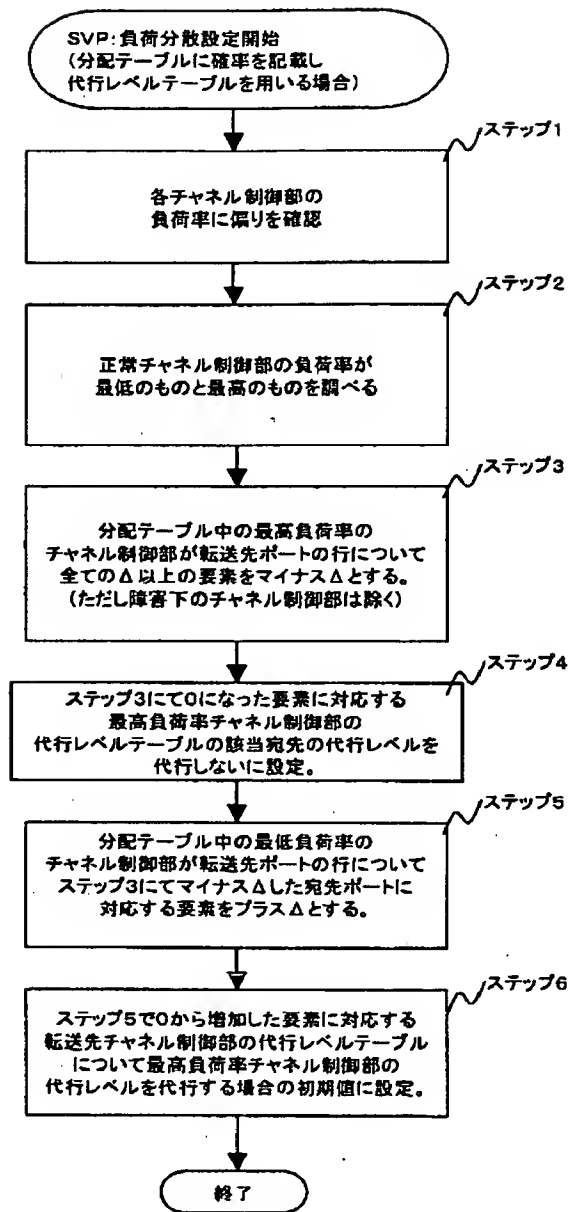
【図 2 2】

図 22



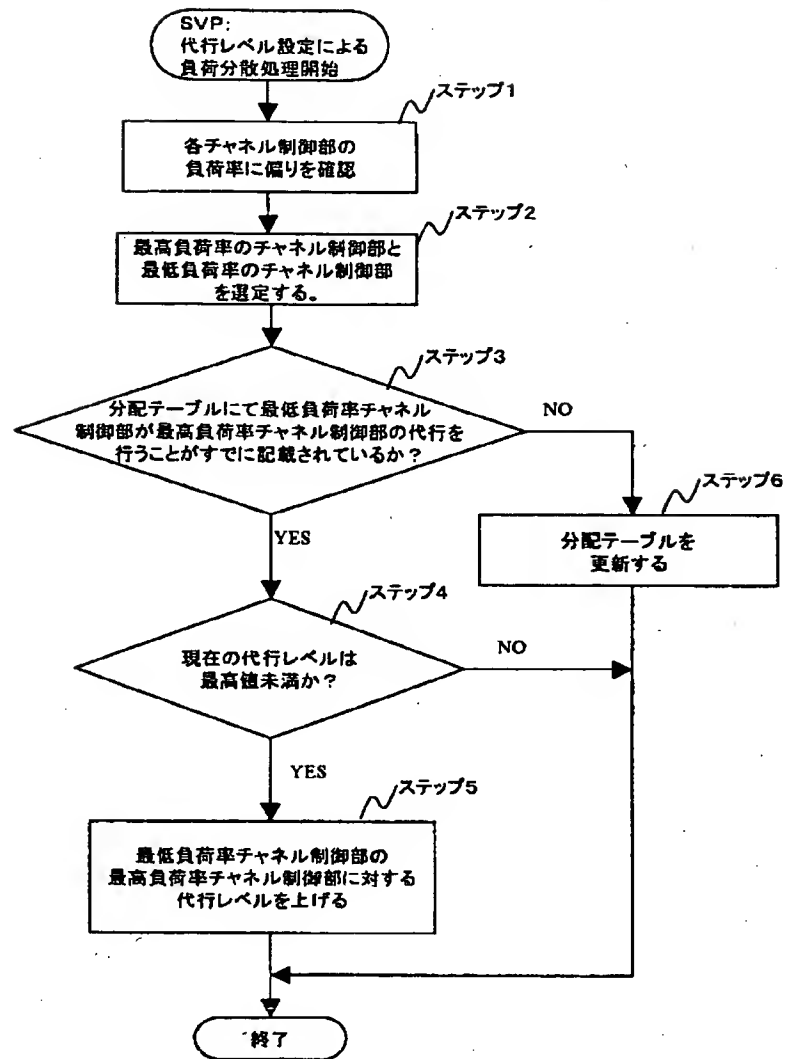
【図 23】

図23



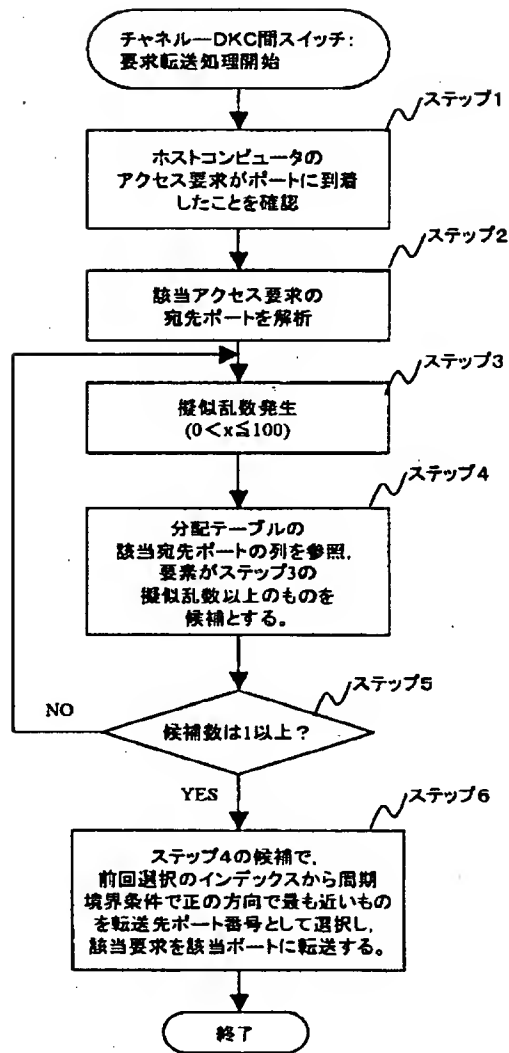
【図 2 4】

図24



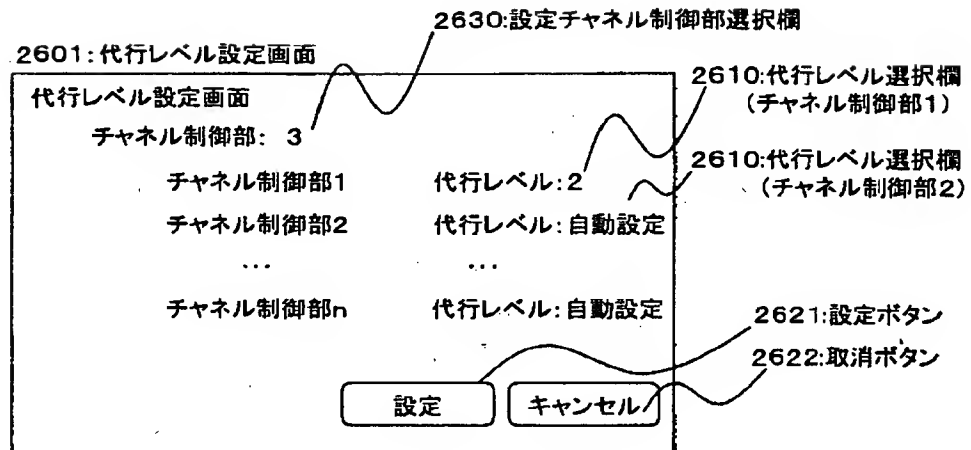
【図 25】

図25



【図 2 6】

図26



【書類名】 要約書

【要約】

【課題】 クラスタ構成のディスクサブシステムにおいて、内部のディスク制御装置間で負荷分散する。

【解決手段】 クラスタ型ディスクサブシステムにおいて、内部にホストコンピュータからの要求の転送先を変更可能とするテーブルを備えたスイッチを有し、高負荷・障害など宛先チャネルの状態に応じてアクセス要求を他のチャネルに転送し、受信したチャネルが代行して要求処理を行うようにした。

【効果】 ホスト側に特別なハードウェア／ソフトウェアを使用することなく、チャネル間やディスク制御装置間の負荷分散やフェイルオーバーが可能である。この結果、ホストコンピュータのアクセス要求が特定のチャネルやディスク制御装置に集中した場合でも性能を引き出すことが可能である。

【選択図】 図1

認定・付加情報

特許出願の番号	特願2002-002936
受付番号	50200020376
書類名	特許願
担当官	第七担当上席 0096
作成日	平成14年 1月11日

<認定情報・付加情報>

【提出日】	平成14年 1月10日
-------	-------------

出 願 人 履 歴 情 報

識別番号 [000005108]

1. 変更年月日 1990年 8月31日

[変更理由] 新規登録

住 所 東京都千代田区神田駿河台4丁目6番地
氏 名 株式会社日立製作所